QUARTETS AND PARAMETER RECOVERY FOR THE GENERAL MARKOV MODEL OF SEQUENCE MUTATION

ELIZABETH S. ALLMAN [1] AND JOHN A. RHODES [2*]

ABSTRACT. A quartet method of phylogenetic inference from biological sequence data might use only 4-dimensional marginal arrays of the joint distribution array of bases in aligned sequences to infer 4-taxon tree topologies, and then use these to infer an n-taxon tree. Since such methods have been advocated as one way of dealing with the computational demands of deducing large phylogenies, understanding the theoretical extent to which quartet methods can ensure the appropriateness of a choice of a model of sequence mutation is of interest. Unfortunately, quartet information cannot confirm that data is in accord with the general Markov model on a tree relating n taxa. However, it can confirm, under certain technical conditions, that there exists not only a unique tree but also unique model parameters consistent with all quartet data. These technical conditions can be phrased in terms of polynomial equalities (phylogenetic invariants) and inequalities in the entries of the marginal arrays.

1. INTRODUCTION

Methods of inference of the evolutionary history leading to currently extant species, or *taxa*, have been transformed in recent years by the ready availability of biological sequence data such as that from DNA. While many approaches to this inference problem have been developed, some of the methods most appealing theoretically are so computationally-intensive that they cannot be carried out exactly when studying a large number of taxa.

One approach to this issue is to first infer phylogenetic trees for smaller subsets of the taxa, and then attempt to combine these smaller trees into a single larger one. In particular, *quartet methods* of phylogenetic inference from biological sequence data for n taxa entail first inferring the topology, perhaps with a measure of confidence, of some or all of the trees relating subsets of 4 taxa, using information on those 4 taxa alone. These quartet trees are then pieced together to form a larger tree, by any one of a number of methods that have been proposed, such as those in [SvH96, BDCG$^+$98, BJK$^+$99, ESSW99] to name only a few.

If the inference problem begins with a collection of aligned sequences, of DNA for instance, from the n taxa, then use of a quartet method would mean using these aligned sequences only in subcollections of 4 at a time. Thus some information is potentially being ignored. Understanding what information may be lost is therefore

Date: May 15, 2004.
1991 *Mathematics Subject Classification.* Primary 92D15; Secondary 60J20 14J99.
Key words and phrases. Quartet, Phylogenetic invariants, Sequence Evolution.
1. Department of Mathematics and Statistics, University of Southern Maine, 96 Falmouth St., Portland, ME, 04104; Tel.: 207-780-4930; Fax: 207-780-5607; Email: eallman@maine.edu .
2*. Corresponding author: Department of Mathematics, Bates College, 3 Andrews Rd., Lewiston, ME, 04240; Tel.:207-786-6403; Fax:207-786-8331; Email: jrhodes@bates.edu .

of interest. (Of course, if for each quartet of taxa we have aligned sequences from different parts of the genome, there may be no loss of information.)

Suppose we restrict ourselves to phylogenetic inference methods that take as input only data on the joint distribution of bases in n aligned sequences from the leaves of the unknown tree. In particular, as with all methods currently in widespread use, we use no information on the location of sites at which observed patterns occur. Then when $n > 4$ no quartet method can fully answer all questions one might have about the relationship of observed pattern frequency data to models of sequence mutation.

To see this, for concreteness let T_r be a rooted bifurcating tree with leaves labeled by the five taxa a, b, c, d, and e, and consider any 2-state model M of sequence mutation. For some particular choice of model parameters $\mathcal{M}$, the expected pattern frequency array (or joint distribution) at the leaves $E_{abcde} = E_{abcde}(T_r, M, \mathcal{M})$, will be a $2\times2\times2\times2\times2$ array, which we believe to be well approximated by the observed joint distribution. Then a quartet method of phylogenetic inference would, by definition, use only the (approximations from the data of the) five 4-dimensional marginal arrays $E_{\Sigma bcde} = \sum_{i=1}^{2} E_{abcde}(i,\cdot,\cdot,\cdot,\cdot)$, $E_{a\Sigma cde}$, $E_{ab\Sigma de}$, $E_{abc\Sigma e}$, and $E_{abcd\Sigma}$ obtained by summing E_{abcde} over one of its indices.

However, if we define the 'checkerboard' array of the same size as E_{abcde} by

$$C(i,j,k,l,m) = (-1)^{i+j+k+l+m},$$

then C has the property that all of its 4-dimensional marginal arrays are identically zero. Thus for any choice of ϵ, $F_{abcde} = E_{abcde}+\epsilon C$ will have the same 4-dimensional marginal arrays as E_{abcde}.

If F_{abcde} were to describe the observed joint distribution of bases in some aligned sequences, then one of two unfortunate possibilities must occur. The first is that F_{abcde} is indeed the expected frequency array for some choice of model parameters, and so quartet information even on 'perfect data' is unable to distinguish between the parameters leading to E_{abcde} and those leading to F_{abcde}. That is, parameters are not identifiable for the model under consideration.

The second possibility is that F_{abcde} is not the expected frequency array for any choice of model parameters, and so while the quartet information might lead us to conjecture we have a distribution exactly consistent with the model, in fact this is not the case.

This situation is analogous to other well-known issues in phylogenetic inference. The first possibility echoes the fact that comparison of sequences for two taxa at a time is not sufficient to identify model parameters under many simple models (e.g., general Markov and submodels). Note that by Chang's work [Cha96], however, for the general Markov model 3-taxon comparisons are sufficient to determine parameters under mild technical assumptions, and so the first possibility can be ruled out for that model. Still, for a more general model incorporating rate variation or dependencies among sites this is not necessarily the case.

The second possibility is reminiscent of an issue that arises in using distance methods for phylogenetic inference. One might assume a certain model describes the evolutionary process leading to some sequence data, and using an appropriate model-based distance formula, calculate distances between terminal taxa. Then these distances may be reasonably consistent with a certain metric tree. Nonetheless, the full pattern frequency data may still be inconsistent with the model underlying the distance formula used. In other words, the work done in verifying that

the distances exactly fit a tree does not verify that the model fits the full data, or that the distance formula used was an appropriate one.

While quartet methods, then, cannot verify the full fit of a model to data, the computational tasks presented by phylogenetic inference for n taxa when n is large still make them attractive. Thus one goal of this paper is an understanding of the extent to which a quartet method, or more generally a k-tet method, can ensure model fit. We of course are focusing on what can in principal be rigorously verified by examining quartets, and not on the more statistical issue of how to deal with the stochastic variation which will make real data fail to exactly pass any such test.

By restricting our consideration to the general Markov model of base substitution, we can obtain some results which are less pessimistic than the preceding discussion. Note that the general Markov model includes as special cases all those commonly considered in the current literature — with the important exception of those allowing rate variation across sites.

While quartet information is not sufficient for verifying that there are model parameters for the full n-taxon tree consistent with the full n-taxon data, it is sufficient (under certain technical assumptions) for ensuring there are model parameters for the full n-taxon tree that are consistent with all the quartet data. To be more precise, fix an n-taxon tree T and suppose we wish to test whether a particular n-dimensional array X is a joint distribution array describing pattern frequencies for the tree T and some parameter choice $\mathcal{M}$. Assume that for the induced tree T_Q associated to each quartet Q we can find parameters $\mathcal{M}_Q$ that would produce the quartet distribution arising from X. Then Theorem 5 shows that, under mild technical conditions, we can indeed find parameters $\mathcal{M}$ for the entire n-taxon tree that will produce all the same quartet distributions. While it is still possible that $X \neq E(\mathcal{M})$, Proposition 11, which generalizes the checkerboard example above, then characterizes how they may differ.

To put this in perspective, the situation is much different if one investigates the same issue for triads (3-taxon subsets) instead of quartets. Recall that in [AR03], Section 7 it was pointed out that one could have a 4-dimensional array, all of whose 3-dimensional marginal arrays satisfied all phylogenetic invariants for the 3-taxon tree, yet which failed to satisfy the 4-taxon invariants. Since [AR03] also showed that satisfying such invariants is roughly equivalent to arising from a parameter choice, this means one would suspect there are examples of 4-dimensional arrays that are not joint distribution arrays for the general Markov model, yet all of whose 3-dimensional marginal arrays are. This suspicion is in fact correct; we construct a specific example of this in Section 7.

Chang's important result — that for the general Markov model triad information is enough for identifying parameters from the joint distribution array for an n-taxon tree — of course assumes one has a joint distribution array for the model to begin with. Our example shows that if one does not know whether an array really arises from the model, then one may be able to identify 'local' (triad) parameters though no 'global' (n-taxon) parameters exist that are consistent with them. Quartet considerations, however, can ensure the existence of global parameters, as Theorem 5 shows.

One might compare this result to the simpler issue of using scalar distances between leaves to determine a tree. As is well known, any distance data can fit a 3-taxon tree (allowing negative lengths), while for four or more taxa this is not

the case. That 4-point conditions of the sort introduced by Buneman [Bun71] be satisfied on all quartets is both necessary and sufficient for distance data to fit a tree, regardless of the number of taxa. Thus our results can be viewed as an analog of the 4-point condition for parameters of the general Markov model, rather than for distances.

The results of [AR03] show how one can ensure an array X is an expected frequency array and has identifiable model parameters by requiring it to satisfy certain explicitly-given phylogenetic invariants and be 'near diagonal'. This allows us to replace the unwieldy assumption of Theorem 5 that parameters satisfying certain technical conditions be recoverable for every quartet, with conditions that the n-dimensional array be near diagonal and satisfy certain polynomial conditions induced from 4-taxon subtrees. However, since phylogenetic invariants are insensitive to the difference between stochastic parameters and more general complex parameters, one only gets conditions ensuring global complex parameters. In fact, this feature of invariants leads us to take extra care to phrase all earlier results in the paper in a form appropriate to complex parameters.

The explicit set of invariants of [AR03] was in fact constructed for κ-base sequences on the n-taxon tree. While the degree of the polynomials was bounded by $\kappa+1$, the cardinality of that set grows exponentially with n. On the other hand, the cardinality of the invariants arising from quartets grows only polynomially with n. Thus a much smaller set of invariants than discussed in [AR03] can be used to test not for full model fit, but at least for the existence of a set of parameters consistent with all quartet data.

Finally, we conclude with a restatement of some of our results in more algebraic-geometric language. While this perspective de-emphasizes certain aspects of the results, it may be a helpful one for further developments.

2. Notation and Terminology

By an *n-taxon tree* T we mean an unrooted tree with n leaves, labeled by the taxa $a_1, \ldots, a_n$, and with all internal vertices of valence 3. Note that internal vertices are not labeled, though we will occasionally use designations such as v or w for them, and x or y for arbitrary vertices. A *rooted* n-taxon tree is a pair (T, r), where r is any choice of a vertex in T, either internal or a leaf. Our usage thus requires that if a root r is internal to the tree, it has valence 3, rather than the more common assumption of valence 2. In the context of the general Markov model, where roots are essentially arbitrarily chosen, we lose little generality by this and gain some simplifications.

We denote an undirected edges between vertices x and y by $x \leftrightarrow y$, and directed edges by $x \rightarrow y$. The edges of a rooted tree will generally be directed away from the root.

Two taxa a_i and a_j are said to be *neighbors* in a tree T if the leaves they label are adjacent to a common vertex — that is, if for some vertex v there exist edges $a_i \leftrightarrow v$ and $a_j \leftrightarrow v$.

Throughout, κ denotes a fixed positive integer, which is interpreted as the number of bases, or letters, from which sequences are composed. Thus $\kappa = 4$ is appropriate for describing DNA sequences.

Definition. For a rooted tree (T, r), *parameters* $\mathcal{M}_r = (\mathbf{p}_r; M_{xy})$ for the general Markov model with κ bases is a collection of a row vector $\mathbf{p}_r \in \mathbb{C}^\kappa$ and matrices $M_{xy} \in M_{\kappa\times\kappa}(\mathbb{C})$ for each edge $x \to y$ in T directed away from r, with the properties that $\mathbf{p}_r\mathbf{1} = 1$ and $M_{xy}\mathbf{1} = \mathbf{1}$, where $\mathbf{1} = (1, 1, \ldots, 1)^T$. We refer to $\mathbf{p}_r$ as the *root distribution vector* and each M_{xy} as a *Markov matrix*. If all the entries of $\mathbf{p}_r$ and M_{xy} are real and non-negative, then we say the parameters $\mathcal{M}$ are *stochastic*.

Given stochastic parameters $\mathcal{M}_r$ for a tree (T, r), we interpret the root distribution vector $\mathbf{p}_r$ as having entries giving the frequencies of various bases in a sequence at r. The Markov matrices M_{xy} have entries giving conditional probabilities of various base substitutions along the edge $x \to y$. This leads to polynomial expressions (in terms of the scalar entries in the parameters) for the expected frequencies of various patterns of bases at the leaves of T. These can be organized into an n-dimensional $\kappa \times \cdots \times \kappa$ array $E(T, \mathcal{M}_r) = E_{a_1\ldots a_n}(\mathcal{M}_r) = E(\mathcal{M}_r)$, where the $(i_1, i_2, \ldots, i_n)$ entry is the expected frequency of the pattern with base i_k at taxon a_k. Thus $E(\mathcal{M}_r)$ gives the joint distribution of bases at the leaves of T.

Even when the parameters are not stochastic, we use $E(\mathcal{M}_r)$ to denote the array obtained from $\mathcal{M}_r$ by the same polynomial expressions as described above, as in [AR03].

Definition. Suppose T is an n-taxon tree. If $K \subseteq \{a_1, a_2, \ldots, a_n\}$ is a k-element subset of the taxa on T, we refer to K as a *k-tet*. For $k = 3, 4$, we use the terms *triad* and *quartet*, respectively. Associated to a k-tet K is the k-taxon tree T_K induced from T with leaves labeled by K. The vertices of T_K are a subset of the vertices of T, while the edges of T_K correspond to paths of length ≥ 1 in T.

Definition. If for a path $x_0 \to x_1 \to \cdots \to x_m$ in T, matrices $M_{x_i x_{i+1}}$ have been defined, then by $M_{x_0 x_m}$ we mean the product $M_{x_0 x_1} M_{x_1 x_2} \cdots M_{x_{m-1} x_m}$.

Note that the last definition allows us to pass naturally from a set of parameters $\mathcal{M}_r$ for (T, r) to an *induced* set of parameters $\mathcal{M}_{r,K}$ for (T_K, r), as long as T_K has r as one of its vertices.

Definition. By a *pattern frequency array* for the taxa $\{a_1, a_2, \ldots, a_n\}$, we mean any n-dimensional $\kappa \times \cdots \times \kappa$ array $X = X_{a_1 a_2 \ldots a_n} \in \mathbb{C}^{\kappa^n}$. If the entries are real, non-negative, and sum to 1, then we say the array is *stochastic*.

Note that while $E(T, \mathcal{M}_r)$ is an example of a stochastic pattern frequency array for the taxa $\{a_1, a_2, \ldots, a_n\}$, most stochastic pattern frequency arrays are not of the form $E(T, \mathcal{M}_r)$ for any choice of T and $\mathcal{M}_r$. In fact, gaining a good understanding of what arrays X are of this form is a central issue.

Definition. Suppose X is a pattern frequency array for $\{a_1, a_2, \ldots, a_n\}$ and $K = \{a_{i_1}, \cdots, a_{i_k}\}$ is a k-tet, where $i_1 < i_2 < \cdots < i_k$. Then the marginal array of X of dimension k obtained by summing over the indices i_j for $a_{i_j} \notin K$ is denoted by

$$X_K = X_{a_{i_1} a_{i_2} \ldots a_{i_k}} = X_{\Sigma\ldots\Sigma a_{i_1}\Sigma\ldots\Sigma a_{i_2}\Sigma\ldots\ldots\Sigma a_{i_k}\Sigma\ldots\Sigma}.$$

X_K is the pattern frequency array for the k-tet K *induced* from X.

If K is a k-tet with r a common vertex of T and T_K, and $\mathcal{M}_r$ parameters for (T, r), then one readily sees that $E(\mathcal{M}_r)_K = E(\mathcal{M}_{r,K})$.

In biological circumstances expected frequency arrays typically have their largest entries on the diagonal. This is because model parameters describe base substitution processes where most sites are left unchanged. This feature is essential in real

data, since most sites must be identical in order to identify related sequences and to align them.

However, if an expected frequency array X has its only non-zero entries on the diagonal, then for any n-taxon (T,r), there is a choice of $\mathcal{M}_r$ with $X = E(T, \mathcal{M}_r)$: For all edges let $M_{xy} = I$, the identity matrix; and let $\mathbf{p}_r$ be composed of the diagonal entries of X. While this is a trivial case for the purposes of phylogenetic inference, since X places no restriction on T, we would like arrays used in inference to be somewhat close to such trivial arrays, since this assumes mutation is rare.

Thus, we define:

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.

We will be interested primarily in arrays X that are *near-diagonal* in the sense that they are close (by the Euclidean metric) to phylogenetically trivial arrays, but that are not themselves phylogenetically trivial.

3. Coherent, Identifiable, and Distance-informative Parameters

As was proved in [SSH94], when dealing with the general Markov model, the choice of a root in a tree is largely irrelevant — usually we can move the root and choose new model parameters without changing the expected pattern frequencies at the leaves. Even with a root specified, though, the map $\mathcal{M}_r \mapsto E(\mathcal{M}_r)$ from model parameters to their expected pattern frequency array is not injective, but rather is typically $(\kappa!)^{n-2}$–to–1 (see [Cha96, AR03]). Finally, for inferring a tree it is convenient if (complex) parameters are well-behaved under the log-det distance.

All of these issues will appear in our efforts to 'piece together' parameters for quartet trees to obtain parameters for a larger tree. We will use the log-det distance to ensure the needed larger tree exists. As not all quartets trees will contain any chosen root for the larger tree, we have to consider different vertices as the root at different times. In addition, the non-injectivity of the expected frequency map will mean that we have to adjust quartet parameters to ensure they 'fit together'.

We therefore define the notions of *coherent, identifiable, and distance-informative parameters* to encapsulate the mild technical conditions we need to overcome these obstacles.

Definition. Suppose T is an unrooted n-taxon tree. Then *coherent parameters* for the general Markov model on T are a collection $\mathcal{M} = (\mathbf{p}_z; M_{xy})$ containing

- for each vertex z of T, a row vector $\mathbf{p}_z$ with all non-zero entries and $\mathbf{p}_z\mathbf{1} = 1$, and
- for each directed edge $x \to y$ in T, a $\kappa \times \kappa$ complex matrix M_{xy} with $M_{xy}\mathbf{1} = \mathbf{1}$,

such that for every edge $x \leftrightarrow y$ of T the following conditions hold:

(i) $\mathbf{p}_y = \mathbf{p}_x M_{xy}$
(ii) $M_{yx} = D_y^{-1} M_{xy}^T D_x$ where $D_x = \mathrm{diag}(\mathbf{p}_x)$.

In addition, if all M_{xy} are non-singular we say the parameters are *identifiable*. If for all pairs u, v of internal vertices,

$$\left|\det\left(D_u^{1/2} M_{uv} D_v^{-1/2}\right)\right| \neq 1,$$

we say the parameters are *distance-informative*. By *c.i.d. parameters* we mean coherent, identifiable, and distance-informative parameters.

Notice that if coherent model parameters $\mathcal{M}$ for an n-taxon tree T are given, then in a natural way $\mathcal{M}$ induces coherent model parameters $\mathcal{M}_K$ on any k-tet tree T_K. The base distributions at vertices in T_K are inherited from $\mathcal{M}$. Moreover, if a directed edge $x \to y$ of T_K corresponds to a path, $x = x_0 \to x_1 \to \cdots \to x_m = y$, in T, then we have $M_{xy} = M_{x_0x_1}M_{x_1x_2}\cdots M_{x_{m-1}x_m}$ in $\mathcal{M}_K$. It is straight-forward to see that this defines coherent parameters for T_K. Moreover, c.i.d. parameters on T induce c.i.d. parameters on T_K.

Notice also that if c.i.d. parameters $\mathcal{M}$ for T are given, and r is any choice of a vertex of T as a root, then we can induce parameters $\mathcal{M}_r$ for the rooted tree (T, r) by taking from $\mathcal{M}$ the vector $\mathbf{p}_r$ together with M_{xy} for each edge $x \to y$ in T directed away from r. These Markov parameters will have the special features that

(1) no entry of $\mathbf{p}_r$ is zero,
(2) if $\mathcal{M}_r$ is used to compute the base distribution $\mathbf{p}_x = \mathbf{p}_r M_{rx}$ at any node x of T, then $\mathbf{p}_x$ will have all non-zero entries,
(3) each M_{xy} is non-singular,
(4) for any pair u, v of internal nodes, $\left|\det\left(D_u^{-1/2}E_{uv}(\mathcal{M}_r)D_v^{-1/2}\right)\right| \neq 1$, where $E_{uv}(\mathcal{M}_r)$ denotes the expected frequency array of bases at u and v.

Conversely,

Lemma 1. *If parameters $\mathcal{M}_r$ for a rooted tree (T, r) satisfy the conditions 1 and 2 above, then there exist unique coherent parameters $\mathcal{M}$ on T which induce $\mathcal{M}_r$. If, in addition, conditions 3 or 4 are satisfied, $\mathcal{M}$ will be identifiable or distance-informative, respectively.*

Proof. Let y be any vertex in T and $r = x_0 \to x_1 \to \cdots \to x_m = y$ the path from the root r to y in (T, r). Define the base distribution at y by $\mathbf{p}_y = \mathbf{p}_r M_{ry} = \mathbf{p}_r M_{x_0x_1}\cdots M_{x_{m-1}x_m}$, so $\mathbf{p}_y$ has non-zero entries and $\mathbf{p}_y\mathbf{1} = 1$. For each directed edge $x \to y$ in (T, r), define $M_{yx} = D_y^{-1}M_{xy}^T D_x$, so $M_{yx}\mathbf{1} = \mathbf{1}$. Also,

$$\mathbf{p}_y M_{yx} = \mathbf{p}_y D_y^{-1} M_{xy}^T D_x = (1, \cdots, 1)M_{xy}^T D_x = (1, \cdots, 1)D_x = \mathbf{p}_x.$$

Uniqueness of $\mathcal{M}$ is clear.

If condition 3 holds, M_{yx} will be non-singular, so $\mathcal{M}$ is identifiable. Finally, if condition 4 holds, then since for any pair u, v of internal vertices,

$$\left|\det\left(D_u^{1/2}M_{uv}D_v^{-1/2}\right)\right| = \left|\det\left(D_u^{-1/2}E_{uv}(\mathcal{M}_r)D_v^{-1/2}\right)\right| \neq 1,$$

$\mathcal{M}$ will be distance-informative. □

Remark. If $\mathcal{M}_r$ is a stochastic set of model parameters, as in any biologically meaningful model, then condition 2 is implied by conditions 1 and 3. In this situation, since all the M_{xy} are invertible, no column of any M_{xy} is identically zero. Since the base distribution $\mathbf{p}_r$ has positive entries, and each column of M_{ry} has non-negative entries with at least one entry positive, then $\mathbf{p}_y = \mathbf{p}_r M_{ry}$ has positive real entries.

Also, for stochastic parameters condition 4 is equivalent to a requirement that $D_u^{-1}E_{uv}(\mathcal{M}_r) = M_{uv}$ not be a permutation matrix, or equivalently that its determinant not be ± 1. For in [Ste94], it is shown that if $\det(D_u^{-1}E_{uv}(\mathcal{M}_r)) \neq \pm 1$ then

$\left|\det\left(D_u^{-1/2}E_{uv}(\mathcal{M}_r)D_v^{-1/2}\right)\right| < 1$, while if $D_u^{-1}E_{uv}(\mathcal{M}_r) = P$ is a permutation, then $\left|\det\left(D_u^{-1/2}E_{uv}(\mathcal{M}_r)D_v^{-1/2}\right)\right| = \left|\det\left(D_u^{1/2}PD_v^{-1/2}\right)\right| = \det\left(D_uD_v^{-1}\right)^{1/2}$, but since $\mathbf{p}_v = \mathbf{p}_uP$, this last quantity is 1.

Proposition 2. *Let coherent parameters $\mathcal{M}$ for T be given. For any choice of a vertex r as a root for T, let $\mathcal{M}_r$ be the parameters for (T, r) induced from $\mathcal{M}$. Then $E(M_r)$, the expected frequency array of patterns at the leaves of T, is independent of the choice of r.*

Proof. This is a straight-forward modification of Theorem 2 of [SSH94]. □

We therefore denote by $E(\mathcal{M})$ the common values of $E(\mathcal{M}_r)$ for any choice of a root r. One then checks that $E(\mathcal{M})_K = E(\mathcal{M}_K)$.

As noted in [Cha96, AR03], there is ambiguity in identifying parameters $\mathcal{M}$ for a tree from an array $E(\mathcal{M})$ of expected pattern frequencies at the leaves of a tree. Since the expected pattern frequency array does not track what base occurs at a site at any internal node — these represent 'hidden variables' in our model — certain permutations of the rows and columns of the Markov matrix parameters in $\mathcal{M}$ may be applied without affecting the entries of $E(\mathcal{M})$ (See Proposition 3 of [AR03] for details, or [Cha96]). Indeed, the identification of parameters is based on the calculation of the eigenvectors and eigenvalues of certain matrices, yet there is no natural order on the eigenvectors. To address this ambiguity, we start with the following definition.

Definition. Let (T, r) be a rooted n-taxon tree with leaves $a_1, \cdots, a_n$, internal nodes $v_1, \cdots, v_{n-2}$, and suppose $\mathcal{M}_r = (\mathbf{p}_r; M_{xy})$ is some choice of model parameters. Let $\sigma = (P_{v_1}, \cdots, P_{v_{n-2}})$, be a choice of $\kappa \times \kappa$ permutation matrices, one for each internal vertex of T. Define $P_{a_i} = I$, the identity matrix, for each leaf a_i.

Then the model parameters $\mathcal{M}_r^\sigma = (\mathbf{p}_rP_r^T; P_xM_{xy}P_y^T)$ are said to be a *permutation of $\mathcal{M}_r$ at the internal nodes of T*.

Notice further that if $\mathcal{M}$ are coherent parameters for T, then it is also possible to define a permutation $\mathcal{M}^\sigma$ of $\mathcal{M}$. Specifically,

Definition. Let T be an n-taxon tree and $\mathcal{M}$ and $\mathcal{M}'$ coherent model parameters for T. Then $\mathcal{M}'$ is a *permutation of* $\mathcal{M}$, if for some choice of vertex r of T as a root, $\mathcal{M}'_r$ is a permutation of $\mathcal{M}_r$.

The choice of root r in the definition is unimportant; if the condition holds for one choice, it holds for all. Indeed, a permutation $\mathcal{M}^\sigma$ of coherent parameters $\mathcal{M}$ can be obtained by the association of a permutation matrix to each internal node of T and carrying out the appropriate matrix algebra. Also, for c.i.d. parameters $\mathcal{M}$, $\mathcal{M}^\sigma$ is also a collection of c.i.d. parameters.

One readily sees that $E(\mathcal{M}_r) = E(\mathcal{M}_r^\sigma)$ for any σ, and thus $E(\mathcal{M}) = E(\mathcal{M}^\sigma)$. Since in the definition above there is a permutation matrix for each of the internal nodes of the tree T, this means there are up to $(\kappa!)^{n-2}$ permutations of $\mathcal{M}$, each producing the same expected pattern frequency array.

Given c.i.d. parameters for a tree T, then to each edge $x \leftrightarrow y$ of T we can associate the number

$$d_{LD}(x, y) = -\log\left|\det\left(D_x^{1/2}M_{xy}D_y^{-1/2}\right)\right| \in \mathbb{R} \smallsetminus \{0\}, \tag{1}$$

which we view as a generalized edge length. For a path $a_i = x_0 \to \cdots \to x_n = a_j$ between taxa, we are led to the tree distance

$$
\begin{aligned}
(2) \qquad d_{LD}(a_i, a_j) &= \sum_{l=1}^{n} d_{LD}(x_{l-1}, x_l) \\
&= -\log \left| \det \left(D_{a_i}^{1/2} M_{a_i a_j} D_{a_j}^{-1/2} \right) \right| \\
&= -\log |\det(D_{a_i} M_{a_i a_j})| + \frac{1}{2} \log \left| \det(D_{a_i} D_{a_j}) \right| \\
&= -\log |\det(E_{a_i a_j}(\mathcal{M}))| + \frac{1}{2} \log \left| \prod_{m,n=1}^{\kappa} \mathbf{p}_{a_i}(m) \mathbf{p}_{a_j}(n) \right|,
\end{aligned}
$$

the formula for the usual log-det distance. Note, however, that for non-stochastic parameters, our edge and path distances need not be positive, though for paths between internal vertices they must be non-zero. If c.i.d. parameters are stochastic then, by [Ste94] as discussed earlier, these distances are all positive.

Proposition 3. *Let $\mathcal{M}$ and $\mathcal{M}'$ be c.i.d. parameters for n-taxon trees T and T'. If $E(\mathcal{M}) = E(\mathcal{M}')$, then $T = T'$ and $\mathcal{M}' = \mathcal{M}^\sigma$ for some choice σ of permutations at the internal nodes of T.*

Proof. This is essentially proved in [Cha96], though he assumes parameters are stochastic, and rules out permutations at internal nodes by making additional assumptions on the Markov matrices. To extend the proof to the current setting of complex parameters, using the distance-informative assumption, the log-det distance of equation (2) can be used along with the 4-point condition, as formulated to allow real edge lengths by Bandelt and Steel [BS95], to first show the tree topology is uniquely determined. The identifiability assumption then allows one to modify the rest of the proof to show uniqueness of the parameters. A 3-taxon statement in the complex setting appears as Proposition 4 of [AR03]. □

As we piece together parameters for various trees, it will be convenient to use the following terminology.

Definition. Suppose T, T' are n-, n'-taxon trees with coherent parameters $\mathcal{M}$, $\mathcal{M}'$, respectively, and that K is the set of taxa common to T and T'. Then we say $(T, \mathcal{M})$ and $(T', \mathcal{M}')$ are *compatible* when $T_K{=}T'_K$ and $\mathcal{M}^\sigma_K = \mathcal{M}'_K$ for some collection σ of permutations at the internal nodes of T.

Note that by Proposition 3 if K is the set of common taxa for T and T', then for c.i.d. parameters, $(T, \mathcal{M})$ and $(T', \mathcal{M}')$ are compatible if, and only if, $E(\mathcal{M})_K = E(\mathcal{M}')_K$.

4. Quartets and Compatible Parameter Recovery

In this section, we give a set of conditions sufficient for building c.i.d. parameters for a tree from those for smaller trees.

We will use the following terminology:

Definition. A pattern frequency array X for taxa $\{a_1, \ldots, a_n\}$ has *Property (Pk)* if for every k-tet K of taxa, there exists a k-taxon tree T_K and c.i.d. parameters $\mathcal{M}_K$ for T_K such that $X_K = E(\mathcal{M}_K)$.

By Proposition 3, both T_K and $\mathcal{M}_K$ whose existence (Pk) implies are uniquely determined by X.

Notice that (Pk) implies (Pj) for $j < k$, for if c.i.d. parameters $\mathcal{M}_K$ exist for a k-tet K, then they induce c.i.d. parameters $\mathcal{M}_J$ for any $J \subset K$ so that $X_J = (X_K)_J = E(\mathcal{M}_K)_J = E(\mathcal{M}_J)$.

As stated, checking whether an array X has Property (Pk) for some k is not easy. In the next section, we will return to this issue, but we first focus on the implications of (P4).

Proposition 4. *Suppose a pattern frequency array X has Property (P4). Then there is a unique n-taxon tree T which induces all the quartet trees T_Q whose existence (P4) asserts.*

Proof. It has already been pointed out that, for each Q, T_Q is uniquely determined. Defining edge lengths on T_Q by the formula of equation (1), then total distances between taxa on T_Q are in agreement with log-det distances computed from X, as discussed above. Thus for all quartets, the log-det distance satisfies the 4-point condition of [BS95], which allows real distances, and so by Theorem 1 of that paper there exists a unique tree T inducing all the T_Q. □

We strengthen this to:

Theorem 5. *Let X be an n-dimensional pattern frequency array with Property (P4). Then there exists a unique n-taxon tree T and c.i.d. parameters $\mathcal{M}$ for T, unique up to permutations at internal nodes, such for all quartets Q, $(T, \mathcal{M})$ is compatible with the $(T_Q, \mathcal{M}_Q)$ whose existence (P4) asserts.*

Furthermore, $\mathcal{M}$ is stochastic if, and only if, all quartet parameters $\mathcal{M}_Q$ whose existence (P4) asserts are stochastic.

This can be rephrased as

Corollary 6. *Let X be an n-dimensional pattern frequency array with Property (P4). Then there exists a unique pattern frequency array X' such that $X' = E(\mathcal{M})$ for c.i.d. parameters $\mathcal{M}$ for an n-taxon tree T and $X'_Q = X_Q$ for all quartets Q.*

This corollary, combined with the results of Section 6, shows the strengths and weaknesses of quartet information for verifying model fit.

Proof of Theorem 5. Note that Property (P4) together with Proposition 3 tells us not only that Property (P3) holds, but also that the parameters whose existence (P3) asserts are unique, up to permutations at the internal node. Since we will use this fact repeatedly, we refer to it simply as (P3!).

Denote the n taxa by $a_1, \ldots, a_n$. By Proposition 4, there exists a unique tree T that induces each of the quartet trees T_Q. We use induction on n to build c.i.d. parameters for T compatible with those for the quartets.

For the case $n = 4$, we need only observe that by Proposition 3 the parameters whose existence are asserted by (P4) are unique, up to permutations at the internal nodes.

Thus we proceed to the inductive step, considering an n-taxon tree T, $n \geq 5$. Note that for any k-tet K, $4 \leq k \leq n-1$, the marginal array X_K inherits Property (P4) from X.

We assume the vertices of T are labeled in such a way that a_1 and a_2 are neighbors, as are a_{n-1} and a_n. Since any n-taxon tree with $n > 3$ has at least two pairs of neighbors, we lose no generality.

Consider the $(n-1)$-tets $K_1 = \{a_2, a_3, \ldots, a_n\}$ and $K_2 = \{a_1, a_2, \ldots, a_{n-1}\}$. Let $\mathcal{N}_{K_1}$ and $\mathcal{N}_{K_2}$ denote the c.i.d. parameters the inductive hypothesis assures us exist for the trees T_{K_1} and T_{K_2}, so $E(\mathcal{N}_{K_i})_Q = X_Q$ for all $Q \subseteq K_i$. Both K_1 and K_2 contain the $(n-2)$-tet $K_1 \cap K_2 = \{a_2, a_3 \ldots, a_{n-1}\}$ and the inductive hypothesis assures us of c.i.d. parameters on $T_{K_1 \cap K_2}$, unique up to permutations at internal nodes. Replacing $\mathcal{N}_{K_i}$ by $\mathcal{N}^{\sigma}_{K_i}$ if necessary, we may assume that for any edge $x \to y$ in $T_{K_1 \cap K_2}$, the parameter M_{xy} in $\mathcal{N}_{K_i}$ is equal to that from $\mathcal{N}_{K_1 \cap K_2}$. Thus we may assume that the parameters $\mathcal{N}_{K_1}$ and $\mathcal{N}_{K_2}$ agree with one another on edges and paths they have in common.

To define parameters for T, take a_2 as the root. Then, since $n > 4$, each directed edge $x \to y$ in (T, a_2) appears in one or both of (T_{K_1}, a_2), (T_{K_2}, a_2). Let M_{xy} be the corresponding matrix parameter in either $\mathcal{N}_{K_1,a_2}$ or $\mathcal{N}_{K_2,a_2}$, since if a parameter is specified in both, it is the same. Let the root distribution vector $\mathbf{p}_{a_2}$ be that which also appears in both $\mathcal{N}_{K_1,a_2}$ and $\mathcal{N}_{K_2,a_2}$. Finally, with $\mathcal{M}_{a_2}$ this collection of parameters for T, let $\mathcal{M}$ be the corresponding c.i.d. parameters.

We will show $\mathcal{M}$ is compatible with all quartet parameters whose existence (P4) implies. Obviously $\mathcal{M}$ is compatible with the $(n-1)$-tet parameters $\mathcal{N}_{K_1}$ and $\mathcal{N}_{K_2}$, which, by induction, are in turn compatible with any quartet parameters for $Q \subset K_1$ or $Q \subset K_2$. Thus there is nothing to show, except for quartets of the form $Q = \{a_1, a_i, a_j, a_n\}$.

Now suppose $Q = \{a_1, a_i, a_j, a_n\}$ is a quartet containing both a_1 and a_n, and let $\widetilde{\mathcal{M}}_Q = (\tilde{\mathbf{p}}_z; \widetilde{M}_{xy})$ be the c.i.d. parameters for T_Q which (P4) ensures exist. We must consider two cases, depending on whether a_1 and a_n are neighbors in T_Q. For both, we temporarily designate a_i as the root.

If a_1 and a_n are not neighbors in T_Q, without loss of generality, assume a_1 and a_i are neighbors joined at v, and a_j and a_n are neighbors are joined at w. Then since both T_Q and T_{K_2} contain the triad $\{a_1, a_i, a_j\}$, (P3!) implies that we can, by applying a permutation at v to $\widetilde{\mathcal{M}}_Q$, assume that $\widetilde{M}_{a_i v} = M_{a_i v}$, $\widetilde{M}_{v a_1} = M_{v a_1}$, and $\widetilde{M}_{v a_j} = M_{v a_j}$. Similar reasoning with T_{K_1}, after possibly applying a permutation at w to $\widetilde{\mathcal{M}}_Q$, gives $\widetilde{M}_{a_i w} = M_{a_i w}$, $\widetilde{M}_{w a_j} = M_{w a_j}$, and $\widetilde{M}_{w a_n} = M_{w a_n}$. Moreover, because $\widetilde{M}_{a_i w} = \widetilde{M}_{a_i v}\widetilde{M}_{vw}$, $M_{a_i w} = M_{a_i v}M_{vw}$, and $\widetilde{M}_{a_i v} = M_{a_i v}$ is invertible, it follows that $\widetilde{M}_{vw} = M_{vw}$. Thus, $\widetilde{\mathcal{M}}_Q$ is compatible with $\mathcal{M}$.

If a_1 and a_n are neighbors in T_Q, let v be the vertex where they are joined, and w the vertex where a_i and a_j are joined. By (P3!) for the triad $\{a_1, a_i, a_j\}$ which lies in $Q \cap K_2$, by applying a permutation at w to $\widetilde{\mathcal{M}}_Q$ if necessary, we may assume that $\widetilde{M}_{a_i w} = M_{a_i w}$ and $\widetilde{M}_{w a_j} = M_{w a_j}$. While we want to show that $\mathcal{M}$ is compatible with $\mathcal{M}_Q$ on the other three edges of T_Q, we cannot do so directly by considering only the triads $Q \cap K_1$ and $Q \cap K_2$, since that gives only that $\widetilde{M}_{w a_1} = M_{w a_1}$ and $\widetilde{M}_{w a_n} = M_{w a_n}$, equalities of products of matrices in $\widetilde{\mathcal{M}}_Q$ and $\mathcal{M}$.

Instead, consider the quartet $Q' = \{a_1, a_2, a_i, a_n\}$ and the coherent model parameters $\widetilde{\mathcal{M}}_{Q'}$ for $T_{Q'}$ that (P4) guarantees. Since a_1 and a_n are not neighbors in $T_{Q'}$, we have already shown that $\widetilde{\mathcal{M}}_{Q'}$ is compatible with $\mathcal{M}$. Thus (applying permutations if necessary) we may assume $\widetilde{\mathcal{M}}_{Q'} = \mathcal{M}_{Q'}$, and therefore denote

the matrix parameters in this set by M_{xy}. However, $Q \cap Q' = \{a_1, a_i, a_n\}$ and so by (P3!), $\widetilde{\mathcal{M}}_Q$ and $\mathcal{M}_{Q'}$ must agree (after a permutation of $\widetilde{\mathcal{M}}_Q$ at v if necessary) on the triad $\{a_1, a_i, a_n\}$. Specifically, this means that $\widetilde{M}_{a_i v} = M_{a_i v}$. So, $\widetilde{M}_{a_i w}\widetilde{M}_{wv} = M_{a_i w}M_{wv}$ and therefore $\widetilde{M}_{wv} = M_{wv}$, since $\widetilde{M}_{a_i w} = M_{a_i w}$ is invertible. It remains only to show that $\widetilde{M}_{va_1} = M_{va_1}$ and $\widetilde{M}_{va_n} = M_{va_n}$, which follows immediately from $\widetilde{M}_{wa_1} = M_{wa_1}$ and $\widetilde{M}_{wa_n} = M_{wa_n}$ now that we have $\widetilde{M}_{vw} = M_{vw}$.

Thus $\mathcal{M}$ is compatible with the parameters for all quartets of T. The uniqueness of $\mathcal{M}$, up to permutation at internal nodes, follows from the fact that $\mathcal{M}$ is determined by all the quartet parameters it induces, and each of these are unique up to permutation at internal nodes. □

5. Phylogenetic Invariants and Quartets

Despite its theoretical interest, if Theorem 5 is to have any practical use, a means of checking whether a pattern frequency array X has Property (P4) is needed. Of course, one could consider each quartet in turn, and through a calculation of eigenvectors and eigenvalues, attempt to compute parameters for the possible quartet trees. However, two objections to this approach might come to mind.

The first is simply that computing eigenvectors is perhaps more work than is absolutely necessary, though for the size of matrices of interest in biological situations the effort does not seem excessive.

The second, more serious, objection is that when working with data, one should not expect an observed pattern frequency array X to *exactly* have Property (P4). One would instead like to test whether X is in some sense close to an array with Property (P4). As one approach to this issue, we will develop polynomial conditions that can imply (P4).

Recall that if (T, r) is an n-taxon tree, then a *phylogenetic invariant* for the general Markov model on (T, r) is a polynomial $p(X) = p(X_{a_1 a_2 \ldots a_n})$ in κ^n variables which vanishes for $X = E(\mathcal{M}_r)$, for any choice of stochastic parameters $\mathcal{M}_r$ on the rooted tree. As discussed in [AR03], this notion is independent of the choice of root r, or whether it is phrased for complex parameters rather than just stochastic ones. One 'obvious' invariant for the general Markov model is the trivial one,

$$p_0(X) = 1 - \sum_{i_1, i_2, \ldots, i_n = 1}^{\kappa} X(i_1, i_2, \ldots, i_n),$$

which merely states that the sum of all expected pattern frequencies must be 1.

If K is a k-tet of the taxa labeling an n-taxon tree T, then any invariant for T_K gives rise to an invariant for T as follows: If $p(X_K)$, a polynomial in κ^k variables, vanishes for all $X_K = E(\mathcal{M}_K)$, then $\tilde{p}(X) = p \circ \mu_K(X)$, where $\mu_K : X \to X_K$ is the marginalization map, will vanish for all $X = E(\mathcal{M})$. More informally, $\tilde{p}(X)$ is obtained from $p(X_K)$ by replacing any variable in $p(X_K)$ with the sum of κ^{n-k} variables, one for each possible choice of base appearing at the $n - k$ taxa not in K. Note that the trivial invariant for T_K gives rise to the trivial invariant for T via the above map.

Several constructions of invariants for the general Markov model were introduced in [AR03]. Of particular note was that method for finding invariants referred to as

commutation relations. The invariants constructed by this method were particularly valuable since, under some additional technical assumptions, one could even deduce that an array X satisfying them was of the form $X = E(\mathcal{M})$ for a tree T. These will therefore play a role in finding polynomial conditions to imply (P4).

For the 4-taxon tree T_0 with neighbor pairs of taxa a, b and c, d, we explicitly give the invariants we will need. Defining several matrices as marginal and cross-sectional arrays of X by

$$X_{i\Sigma cd}(k,l) = \sum_{j=1}^{\kappa} X_{abcd}(i,j,k,l), \qquad X_{\Sigma\Sigma cd}(k,l) = \sum_{i=1}^{\kappa} X_{i\Sigma cd},$$

$$X_{abij}(k,l) = X_{abcd}(k,l,i,j,), \qquad X_{ab\Sigma\Sigma} = \sum_{i,j=1}^{\kappa} X_{abij},$$

then as shown in [AR03], for all choices of $1 \leq i, j, k, l \leq \kappa$ the entries of the matrices

$$X_{i\Sigma cd}\mathrm{Cof}(X_{\Sigma\Sigma cd})^T X_{j\Sigma cd} - X_{j\Sigma cd}\mathrm{Cof}(X_{\Sigma\Sigma cd})^T X_{i\Sigma cd} \tag{3}$$
$$X_{abij}\mathrm{Cof}(X_{ab\Sigma\Sigma})^T X_{abkl} - X_{abkl}\mathrm{Cof}(X_{ab\Sigma\Sigma})^T X_{abij}$$

are phylogenetic invariants for T_0. Let $\mathcal{S}'(T_0)$ denote the set of all these invariants, together with the trivial invariant for T_0. All of its elements, except the trivial invariant, have degree $\kappa + 1$.

In [AR03], a set of invariants $\mathcal{S}'(T)$ is defined more generally for any n-taxon tree, via induction on n; since we will not use those invariants here, we omit the full definition.

Recall that a set of invariants for T is said to be *parameter-strong* on a set $\mathcal{O}$ if for any $X \in \mathcal{O}$ on which all the invariants vanish, we have $X = E(\mathcal{M}_r)$ for some choice of parameters $\mathcal{M}_r$. The main ingredient for the proof of Theorem 10 below is the following result, used only in the case of $n = 4$ taxa for $T = T_0$.

Theorem 7. *(Theorem 13 of* [AR03]*) Let T be an n-taxon tree with taxa $a_1, a_2, \ldots, a_n$. There exists an open set $\mathcal{O} \subseteq \mathbb{C}^{\kappa^n}$ which contains all phylogenetically trivial arrays and on which $\mathcal{S}'(T)$ is parameter-strong, regardless of which leaf is taken as the root. Moreover, for a fixed choice of root, if $X \in \mathcal{O}$ and $X = E_{a_1\ldots a_n}(\mathcal{M}_r) = E_{a_1\ldots a_n}(\mathcal{M}'_r)$, then $\mathcal{M}'_r = \mathcal{M}^{\sigma}_r$ for some choice σ of permutations at the internal nodes of T_0.*

We need a slight strengthening of this, giving coherent identifiable parameters.

Proposition 8. *Let T be an n-taxon tree, with taxa $a_1, a_2, \ldots, a_n$. There exists an open set $\mathcal{O} \subseteq \mathbb{C}^{\kappa^n}$ which contains all phylogenetically trivial arrays and on which $\mathcal{S}'(T)$ is coherent-identifiable-parameter-strong. That is, if all polynomials in $\mathcal{S}'(T)$ vanish at $X \in \mathcal{O}$, then $X = E_{a_1\ldots a_n}(\mathcal{M})$ for a collection $\mathcal{M}$ of coherent, identifiable parameters on T. Moreover, $\mathcal{M}$ is unique up to permutations at the internal nodes of T.*

Proof. A careful reading of the proofs of Theorems 11 and 13 in [AR03] shows we need only prove that coherent, identifiable parameters may be recovered for a 3-taxon tree. If T_R is the 3-taxon tree for the triad $R = \{a, b, c\}$ with internal node f and rooted at a, then the proof of Theorem 11 of [AR03] shows that $\mathbf{p}_a$, $\mathbf{p}_b$, $\mathbf{p}_c$, $\mathbf{p}_f$ have all non-zero entries and M_{af}, M_{fb} are invertible, though there is no guarantee

that M_{fc} is invertible. Let $\mathcal{O}_1 \subseteq \mathbb{C}^{\kappa^3}$ be the set whose existence is guaranteed by ([AR03], Theorem 11). Let $\mathcal{O}_2 \subseteq \mathbb{C}^{\kappa^3}$ be the open set defined by $\det(X_{ac}) \neq 0$, and let $\mathcal{O} = \mathcal{O}_1 \cap \mathcal{O}_2$. Then if $X \in \mathcal{O}$ satisfies all the specified triad invariants, since $X_{ac} = D_a M_{af} M_{fc}$ and X_{ac} is invertible, so is M_{fc}. □

To ensure that parameters for T_0 are also distance-informative, we place polynomial inequality restrictions on X: Given coherent parameters for T_0, if u, v denote the internal nodes, then $\det(D_u^{1/2} M_{uv} D_v^{-1/2}) \neq \pm 1$ is equivalent to $d_{LD}(u,v) \neq 0$, which in turn is equivalent to

$$d_{LD}(a,c) + d_{LD}(b,d) - d_{LD}(a,b) - d_{LD}(c,d) \neq 0,$$
$$-\log|\det E_{ac}(\mathcal{M})| - \log|\det E_{bd}(\mathcal{M})| + \log|\det E_{ab}(\mathcal{M})| + \log|\det E_{cd}(\mathcal{M})| \neq 0,$$
$$\text{or } \det(E_{ac}(\mathcal{M})E_{bd}(\mathcal{M})) \neq \pm(\det(E_{ab}(\mathcal{M})E_{cd}(\mathcal{M})))$$

Letting $\mathcal{D} = \{\det(X_{ac}X_{bd}) \pm \det(X_{ab}X_{cd})\}$, we thus have

Lemma 9. *If $X = E(\mathcal{M})$ for coherent parameters $\mathcal{M}$ on T_0, and neither polynomial in $\mathcal{D}$ vanishes at X, then $\mathcal{M}$ is distance-informative.*

Of course, if X is phylogenetically trivial, then all 2-dimensional marginal arrays of X are identical, and so an element of $\mathcal{D}$ does vanish on X, as one should have expected.

Now if $Q = \{a_i, a_j, a_k, a_l\}$ is any quartet, there are three quartet trees with these taxa as labels, which we denote by T_Q^i, $i = 1, 2, 3$. For each i, choosing an identification of the pairs of neighbors in T_Q^i with a, b and c, d in T_0, we obtain from $\mathcal{S}'(T_0)$ a set of invariants $\mathcal{S}(T_Q^i)$ for T_Q^i, and from $\mathcal{D}$ a set $\mathcal{D}(T_Q^i)$. We note that there is some arbitrary choice in this definition, in the choice of which neighbor pair is identified with a, b. Nonetheless, we assume such choices have been made so that our sets are well-defined.

We now formulate a polynomial condition that implies Property (P4):

Theorem 10. *There exists an open set $\mathcal{O} \subseteq \mathbb{C}^{\kappa^n}$ which contains all phylogenetically trivial arrays with the following property: If $X \in \mathcal{O}$ is an n-taxon pattern frequency array and for each quartet Q of the taxa there is an i such that all polynomials in $\mathcal{S}(T_Q^i)$ vanish at X_Q while no polynomial in $\mathcal{D}(T_Q^i)$ vanishes at X_Q, then X has Property (P4).*

Proof. For each of the $3\binom{n}{4}$ possible quartet trees T_Q^i, let $\mathcal{O}_Q^i \subseteq \mathbb{C}^{\kappa^4}$ be the set whose existence is guaranteed by Proposition 8. Then for each Q, $\mathcal{O}_Q = \cap_{i=1}^3 \mathcal{O}_Q^i$ contains all phylogenetically-trivial 4-dimensional arrays.

Let $\mu = \oplus \mu_Q : \mathbb{C}^{\kappa^n} \to \oplus_Q \mathbb{C}^{\kappa^4}$, the sum being taken over all quartets Q, so that $\mu(X) = \oplus_Q X_Q$. Let $\mathcal{O} = \mu^{-1}(\oplus_Q \mathcal{O}_Q)$. Then $\mathcal{O}$ is an open set containing all phylogenetic trivial arrays.

Furthermore, if $X \in \mathcal{O}$ and Q is a quartet such that for some i all polynomials in $S(T_Q^i)$ vanish at X_Q and no polynomial in $\mathcal{D}(T_Q^i)$ vanishes at X_Q, then by Proposition 8 and Lemma 9, there exist c.i.d. parameters $\mathcal{M}_Q$ for T_Q^i with $X_Q = E(\mathcal{M}_Q)$. □

Two points concerning this result are worth noting: first, the open set $\mathcal{O}$ is not explicitly given, and second, the parameters $\mathcal{M}$ may be complex rather than stochastic. In fact, an explicit $\mathcal{O}$ could be given by tracking through the various

proofs, though it probably would be much smaller than is necessary. Currently, polynomial conditions on X that will assure parameter values are stochastic are not known.

It would, of course, be straightforward to generalize the work of this section to give polynomial conditions implying Property (Pk) for other values of k.

6. Frequency arrays agreeing on k-tets

Suppose $E = E(T, M, \mathcal{M})$ is the n-dimensional $\kappa \times \kappa \times \cdots \times \kappa$ array of expected pattern frequencies at the leaves for some rooted tree T, model of mutation M, and parameter choice $\mathcal{M}$. Here M might be any model at all; it could be the general Markov model we deal with elsewhere in this paper, or a more restricted one, or one allowing rate variation across sites, or even dependencies between sites, insertions, and deletions. We merely need that the model and parameter choices determine expected pattern frequencies at the leaves somehow.

For fixed k, with $0 \leq k \leq n$, we are interested in identifying all other arrays F with $F_K = E_K$ for all k-tets K. Sequence data described by such an F would be indistinguishable from sequence data described by E, as long as comparison of sequences from at most k taxa at a time is allowed. Thus to any k-tet method, F and E are identical.

More generally, if X and Y are n-dimensional arrays, then their k-dimensional marginal arrays will be identical if, and only if, the array $Z = X - Y$, obtained by component-wise subtraction, has the property that all its k-dimensional marginal arrays are identically zero.

Proposition 11. *For $0 \leq k \leq n$ and L any field, let V_k^n denote the set of n-dimensional $\kappa \times \kappa \times \cdots \times \kappa$ arrays with entries in L having the property that all k-dimensional marginal arrays are identically zero. Then the dimension of V_k^n as a vector space over L is*

$$\dim V_k^n = \sum_{i=k+1}^{n} \binom{n}{i} (\kappa - 1)^i = \kappa^n - \sum_{j=0}^{k} \binom{n}{j} (\kappa - 1)^j.$$

Proof. This may be proved directly by construction of a basis, or by appealing to the more general result of Theorem 2.6 of Hoşten and Sullivant in [HS02] on arrays with certain marginalizations vanishing.

□

Corollary 12. *For fixed E, k, the set of all F such that $F_K = E_K$ for all k-tets K is $E + V_k^n$. This set has dimension $\dim(V_k^n) \sim \kappa^n$ as $n \to \infty$.*

Remark. A illustrative case for biological applications concerns applying a quartet method to DNA sequence data for 5 taxa, so $k = 4$, $\kappa = 4$, and $n = 5$. The proposition then says $\dim V_4^5 = 3^5$. Thus for each 5-dimensional $\kappa \times \kappa \times \cdots \times \kappa$ array in the 4^5-dimensional space of such arrays, there are 3^5 degrees of freedom in choosing other 5-dimensional arrays that behave identically under a quartet method.

Note also that any array of the form $E(\mathcal{M})$ for stochastic parameters $\mathcal{M}$ will have non-negative entries. If the entries are in fact positive (as is true for most choices of $\mathcal{M}$ for many models), then for $Z \in V_m^n$ sufficiently close to the origin $E(\mathcal{M}) + Z$ will also have all positive entries. Thus even considering the necessity

of non-negativity in biologically meaningful data fails to eliminate these degrees of freedom.

7. Triads and Incompatible Parameters

As sections 4 and 5 showed that quartet information is powerful for ensuring an array can be associated with parameters compatible with all quartets, one might ask if a similar result holds for triads. Specifically, one could question whether given any 4-dimensional array X, all of whose triad marginal arrays arose as joint distribution arrays for pattern frequencies for some parameter choices in the 3-taxon general Markov model, is there a choice of parameters $\mathcal{M}$ for a 4-taxon tree that produces a pattern frequency array $E(\mathcal{M})$ with the same triad marginal arrays as X? Note that while there are 3 topologically distinct 4-taxon trees, included as part of this question is the determination of the correct topology. A counterexample to this will illustrate the weakness of triad information alone.

To construct the counterexample, we use the observation in [AR03] that if $\kappa = 2$, then the ideal of phylogenetic invariants for the 3-taxon tree is generated by the stochastic invariant. Thus any $2 \times 2 \times 2$ array whose entries sum to 1 is on the phylogenetic variety, and is therefore likely to arise from model parameters.

One way to find the desired counterexample is to pick a reasonably random $2 \times 2 \times 2 \times 2$ array whose entries sum to 1. By the preceding paragraph, each of the triad marginal arrays are likely to arise from a parameter choice. These parameters can be computed (as outlined in [Cha96] or [AR03]), and then they are likely to be incompatible with one another. However, while this scheme readily produces arrays with the desired properties, the triad parameters are typically complex. In order to illustrate that the triad parameters can all be stochastic, yet no quartet parameters exist, we give a specific example.

For the 4-taxon tree with neighbor pairs a, b joined at u, and c, d joined at v, consider the Jukes-Cantor parameters $\mathbf{p}_a = (.5, .5)$ and $M = \begin{pmatrix} .9 & .1 \\ .1 & .9 \end{pmatrix}$. Let M be the Markov matrix on all edges of the tree except $v \to d$, where M^2 is used, and E_{abcd} the resulting expected frequency array. Let F_{abcd} be the expected frequency array for Jukes-Cantor parameters with M on all edges except $a \to u$, where M^2 is used. Then let $X = (.5)E_{abcd} + (.5)F_{abcd}$, so X is actually an expected frequency array for a model in which sites fall into two equidistributed classes, with slower or faster mutation occurring on certain branches of the tree depending on the class.

A short computation of the triad arrays $X_{abc} = X_{abc\Sigma}$ and $X_{abd} = X_{ab\Sigma d}$ shows

$$X_{ab1\Sigma} = \begin{pmatrix} .3186 & .0466 \\ .0594 & .0754 \end{pmatrix}, \quad X_{ab2\Sigma} = \begin{pmatrix} .0754 & .0594 \\ .0466 & .3186 \end{pmatrix},$$

and

$$X_{ab\Sigma 1} = \begin{pmatrix} .3058 & .0466 \\ .0594 & .0882 \end{pmatrix}, \quad X_{ab\Sigma 2} = \begin{pmatrix} .0882 & .0594 \\ .0466 & .3058 \end{pmatrix},$$

By computing appropriate eigenvectors one can deduce that $X_{abc} = E(\mathcal{M}')$ where $\mathcal{M}'$ is composed of $\mathbf{p}_a = (.5, .5)$,

$$M_{au} = \begin{pmatrix} .86 & .14 \\ .14 & .86 \end{pmatrix}, \quad M_{ub} = M, \quad M_{uc} = M^2.$$

Similarly, $X_{abd} = E(\mathcal{M}'')$ where $\mathcal{M}''$ is composed of $\mathbf{p}_a = (.5, .5)$,

$$M_{au} \approx \begin{pmatrix} .85777 & .14223 \\ .14223 & .85777 \end{pmatrix},\ M_{ub} \approx \begin{pmatrix} .90249 & .09751 \\ .09751 & .90249 \end{pmatrix},\ M_{ud} \approx \begin{pmatrix} .78622 & .21378 \\ .21378 & .78622 \end{pmatrix}.$$

Although these last parameter values are approximate, one can rigorously verify that they are stochastic. Clearly these two sets of parameters are not compatible with each other, so there can be no 4-taxon parameters compatible with them both. Also note that parameters for the remaining two triads will also be stochastic, as the symmetry of the construction of X ensures that, except for labeling, they will be the same as those already given.

Notice, finally, that this example for $\kappa = 2$ is readily modified to give examples for larger κ; simply embed it 'on the diagonal' of a larger array which is otherwise zero.

8. Geometric viewpoint

This section is primarily expository, reinterpreting earlier results in this paper from a more geometric viewpoint. Although algebraic geometry provides a very natural setting for discussions of phylogenetic invariants and related issues, terminology from that field has not been heavily used in the phylogeny literature. (See [Hag00, HL00] for exceptions.) Therefore the texts [CLO97] and [Har92] are suggested for further background. We also draw attention to the interesting works [GSS03] and [Str83]; although phylogenetic inference is not mentioned as an application in those works, the overlap of methods and goals are worth noting.

For a fixed rooted n-taxon tree (T, r) under the general Markov model, the model parameters $\mathcal{M}_r$ include a root distribution vector and Markov matrices for each edge directed away from the root. Equivalently, since T has $(2n-3)$ edges, the stochastic parameter space $\mathcal{P}$ can be viewed as a subset of $[0,1]^N$, where $N = (\kappa-1) + (2n-3)\kappa(\kappa-1)$. The expected pattern frequencies at the terminal nodes arising from a stochastic parameter choice can be viewed as an element of $[0,1]^M \subset \mathbb{R}^M \subset \mathbb{C}^M$, where $M = \kappa^n$. For the general Markov model there are polynomials giving the map from parameters to pattern frequencies $\Phi_0 : \mathcal{P} \to [0,1]^M$. This map of course extends naturally to a polynomial map $\Phi : \mathbb{C}^N \to \mathbb{C}^M$.

One would like to fully understand $\Phi_0(\mathcal{P})$, the image of the stochastic parameter space under Φ_0. A more tractable, yet still difficult, problem would be to understand fully $\Phi(\mathbb{C}^N)$. From the viewpoint of algebraic geometry, Φ is a parameterization of a dense subset of an algebraic variety, and it is natural to seek an implicit description of it — that is, one can ask for the ideal of polynomials which vanish on $\Phi(\mathbb{C}^N)$. This *phylogenetic ideal* $\mathfrak{A}_T$, is the kernel of the map $\mathbb{C}[X_1, \ldots, X_M] \to \mathbb{C}[P_1, \ldots, P_N]$ defined by the substitution $(X_1, \ldots, X_M) = \Phi(P_1, \ldots, P_N)$, and hence is prime. For the general Markov model, $\mathfrak{A}_T$ is independent of the choice of root for T.

The polynomials in the phylogenetic ideal are of course usually termed *phylogenetic invariants* in the phylogeny literature. An explicit set of polynomials proven to generate the ideal is not currently known for the general Markov model. (However, see [SS04] for recent results on group-based models.)

The *phylogenetic variety* $V(\mathfrak{A}_T)$ is the set in $\mathbb{C}^M$ on which all polynomials in the phylogenetic ideal vanish. It is the smallest algebraic variety that contains $\Phi(\mathbb{C}^N)$

(or even $\Phi_0(\mathcal{P})$), but for $\kappa \geq 2$, $n \geq 3$ is strictly larger than $\Phi(\mathbb{C}^N)$. Since the phylogenetic ideal is prime, the phylogenetic variety is irreducible.

At least one case of a phylogenetic variety for the general Markov model appears as a construction in classical algebraic geometry, as is made clear by an observation in [GSS03]. For a 3-taxon tree, consider a fixed base in a sequence at the central vertex. Then the probabilities of its mutation along an edge leading to a leaf can be specified by an element of the projective space $\mathbb{P}^{\kappa-1}$, and the probabilities of it mutating to produce various patterns at the leaves is thus given by an element of the Segre product $\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1}$. Since there are κ possible bases at the central node, taking a sum of κ points on this Segre variety, weighted by the probabilities of those bases occurring, gives the joint distribution of bases at the leaves. But geometrically, this is just a higher secant variety. Thus a projective version of this phylogenetic variety (dropping the trivial invariant) is that it is the κ-secant variety of the Segre product of 3 copies of $\mathbb{P}^{\kappa-1}$. Although there are few classical results on the generators of the associated ideal, determining them remains a focus of research.

While explicit generators of the full phylogenetic ideal are not known, [AR03] provided new constructions of many polynomials in the ideal, and gave some indication that these capture much of the ideal. More precisely, notions of sets of invariants that were strong and parameter-strong on subsets of $\mathbb{C}^M$ were introduced, and explicit sets of invariants were shown to have these properties on a Euclidean open set $\mathcal{O}$ containing all phylogenetically trivial arrays.

While any ideal generated by phylogenetic invariants defines a variety V with $V \supseteq V(\mathfrak{A}_T)$, for a variety generated by a strong set of invariants on a set $\mathcal{O}$, one has

$$V \cap \mathcal{O} = V(\mathfrak{A}_T) \cap \mathcal{O}. \tag{4}$$

If $\mathcal{O}$ is a Euclidean open set and $V \cap \mathcal{O} \neq \emptyset$, then equation (4) implies that

$$V = V(\mathfrak{A}_T) \cup V'$$

where V' is some variety not containing $V(\mathfrak{A}_T)$.

Thus, the results of [AR03] give an explicit set of polynomials for which the associated variety includes $V(\mathfrak{A}_T)$ as one of its irreducible components, with (possibly) a finite number of additional ones. In the case $\kappa \leq 4$, $n = 3$, any additional components must lie within another explicitly known variety defined by certain determinant conditions.

Given a quartet such as $Q = \{a_1, a_2, a_3, a_4\}$ of taxa from the tree T, the associated marginalization map $\mu_Q : \mathbb{C}^M \to \mathbb{C}^{\kappa^4}$ is (up to a scalar factor) just an orthogonal projection of $\mathbb{C}^M$ onto a certain subspace. One readily sees that $\mu_Q(V(\mathfrak{A}_T)) \subseteq V(\mathfrak{A}_{T_Q})$.

We view a quartet method of phylogenetic inference as one that, rather than taking as its input a data point X of pattern frequencies in $\mathbb{C}^M$, uses only the images $\mu_Q(X)$, for some or all quartets Q. In this light, Section 6 of this paper simply characterizes the kernel of the map $\mu = \oplus_Q \mu_Q : \mathbb{C}^M \to \oplus_Q \mathbb{C}^4$, which described the loss of information inherent in quartet methods.

One can also see that μ is injective when restricted to an explicitly determinable Zariski open subset of $V(\mathfrak{A}_T)$. To this end, consider the maps

$$\mathbb{C}^N \xrightarrow{\Phi} V(\mathfrak{A}_T) \xrightarrow{\mu} \prod_Q V(\mathfrak{A}_{T_Q}),$$

and construct the desired subset as follows: If $X \in V(\mathfrak{A}_T)$, then, assuming certain determinants are non-zero, one can attempt to find parameters $P \in \mathbb{C}^N$ with $X = \Phi(P)$ by the computation of simultaneous eigenvectors and eigenvalues of certain matrices defined from X as in [Cha96]. By [AR03], these matrices must commute since X is on the phylogenetic variety, and so P can be found if even one of the matrices has distinct eigenvalues. Since the conditions that the eigenvalues of a matrix be distinct can be expressed as the non-vanishing of a polynomial in the matrix entries, this means there is a Zariski open subset $\mathcal{O}'$ of $V(\mathfrak{A}_T)$ such that if $X \in \mathcal{O}'$ then $X = \Phi(P)$ for some P.

Now the proof of Proposition 3 tells us that on $\Phi^{-1}(\mathcal{O}')$, a Zariski open subset of $\mathbb{C}^N$, the map Φ fails to be injective only because of the issue of permutation of parameters. Moreover, the method of proof shows that the same statement holds true for $\mu \circ \Phi$. (In fact, it would hold even if quartets were replaced by triads.) Taken together, these facts imply that μ is injective on $\mathcal{O}'$.

So far in this discussion, we have fixed an n-taxon tree T. However, the full phylogenetic inference problem in our setting begins with only the taxa specified. Given $X \in \mathbb{C}^M$, we would hope to use only $\mu(X)$ to determine if a tree T exists for which $X \in V(\mathfrak{A}_T)$.

Now for any quartet Q, there are three possible quartet topologies T^i_Q, $i = 1, 2, 3$, and so a necessary condition for T to exist is that for all Q, there exists an i with $\mu_Q(X) \in V(\mathfrak{A}_{T^i_Q})$. This condition is not sufficient, but from Theorem 10, Corollary 6, and Section 6, we obtain

Theorem 13. *For each quartet Q in a collection of taxa, fix a quartet tree T_Q. Then there is a Euclidean open set $\mathcal{O} \subset \mathbb{C}^{\kappa^n}$ containing all phylogenetically trivial arrays and Zariski open sets $\mathcal{O}_Q \subset \mathbb{C}^{\kappa^4}$ such that, if $X \in \mathcal{O}$ and $\mu(X) \in \prod_Q (V(\mathfrak{A}_{T_Q}) \cap \mathcal{O}_Q)$, then there is a unique tree T and a unique point $X' \in V(\mathfrak{A}_T)$ such that $\mu(X') = \mu(X)$. Moreover $\mu^{-1}(\mu(X)) = X' + V_4^n$.*

Of course if the quartet trees T_Q in this statement are not all induced from some n-taxon tree T, then the stated conditions on X cannot be satisfied, since the conclusion is impossible in that case.

A similar statement in which the Euclidean open set $\mathcal{O}$ is replaced by an explicit Zariski open set is also possible, since $\mathcal{O}$ was introduced to ensure certain matrices are diagonalizable, and that can also be done by requiring that their eigenvalues be distinct, as above. However, the resulting Zariski open set would not contain $\Phi(P)$ for many biologically reasonable values of parameters P. Nonetheless, modifications to this idea, where one requires that certain linear combinations of the matrices have distinct eigenvalues, can remedy that flaw.

Note that the Zariski open sets $\mathcal{O}_Q$ of the theorem only exclude points for which the quartet topology is not distinguishable by the log-det distance, and thus these sets have a quite natural role.

Since it focuses solely on the algebraic varieties, the formulation of our results in Theorem 13 hides any reference to the parameter space that really underlies much

of the proofs. This is not fully desirable, however, since the parameters are not just a tool for the proof, but have biological meaning.

References

[AR03] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.

[BDCG+98] A. Ben-Dor, B. Chor, D. Graur, R. Ophir, and D. Pelleg. From four-taxon trees to phylogenies: the case of mammalian evolution. In *Proc. 2nd Annual Int. Conf. Comp. Mol. Biol. (RECOMB)*, pages 9–19, 1998.

[BJK+99] Vincent Berry, Tao Jiang, Paul Kearney, Ming Li, and Todd Wareham. Quartet cleaning: improved algorithms and simulations. In *Algorithms—ESA '99 (Prague)*, volume 1643 of *Lecture Notes in Comput. Sci.*, pages 313–324. Springer, Berlin, 1999.

[BS95] Hans-Jürgen Bandelt and Michael Anthony Steel. Symmetric matrices representable by weighted trees over a cancellative abelian monoid. *SIAM J. Disc. Math.*, 8:517–525, 1995.

[Bun71] Peter Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences*, pages 387–395, Edinburgh, 1971. Edinburgh University Press.

[Cha96] Joseph T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.

[CLO97] David Cox, John Little, and Donal O'Shea. *Ideals, varieties, and algorithms*. Springer-Verlag, New York, second edition, 1997.

[ESSW99] P.L. Erdös, M. Steel, L. Székely, and T. Warnow. A few logs suffice to build (almost) all trees (I). *Random Struct. Alg.*, 14:153–184, 1999.

[GSS03] Luis David Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of Bayesian networks. `arXiv:math.AG/0301255`, 2003.

[Hag00] Thomas R. Hagedorn. Determining the number and structure of phylogenetic invariants. *Adv. in Appl. Math.*, 24(1):1–21, 2000.

[Har92] Joe Harris. *Algebraic geometry : a first course*. Springer-Verlag, New York, 1992.

[HL00] Thomas R. Hagedorn and Laura F. Landweber. Phylogenetic invariants and geometry. *J. Theor. Biol.*, 205:365–376, 2000.

[HS02] Serkan Hoşten and Seth Sullivant. Gröbner bases and polyhedral geometry of reducible and cyclic models. *J. Combin. Theory Ser. A*, 100(2):277–301, 2002.

[SS04] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. `arXiv:q-bio.PE/0402015`, 2004.

[SSH94] M.A. Steel, L. Székely, and M.D. Hendy. Reconstructing trees from sequences whose sites evolve at variable rates. *J. of Comp. Bio.*, 1(2):153–163, 1994.

[Ste94] M. Steel. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Letters*, 7(2):19–24, 1994.

[Str83] V. Strassen. Rank and optimal computation of generic tensors. *Linear Algebra Appl.*, 52/53:645–685, 1983.

[SvH96] K. Strimmer and A. von Haeseler. Quartet puzzling: a maximum likelihood method for reconstructing tree topologies. *Mol. Biol. & Evol.*, 13:964–969, 1996.

Department of Mathematics and Statistics, University of Southern Maine, Portland, Maine 04104

E-mail address: `eallman@maine.edu`

Department of Mathematics, Bates College, Lewiston, Maine 04240

E-mail address: `jrhodes@bates.edu`