SIAM J. APPL. ALGEBRA GEOMETRY
Vol. 8, No. 1, pp. 114–137

Phylogenomic Models from Tree Symmetries*

Elizabeth S. Allman†, Colby Long‡, and John A. Rhodes†

Abstract. A model of genomic sequence evolution on a species tree should include not only a sequence substitution process, but also a coalescent process, since different sites may evolve on different gene trees due to incomplete lineage sorting. Chifman and Kubatko initiated the study of such models, leading to the development of the SVDquartets methods of species tree inference. A key observation was that symmetries in an ultrametric species tree led to symmetries in the joint distribution of bases at the taxa. In this work, we explore the implications of such symmetry more fully, defining new models incorporating only the symmetries of this distribution, regardless of the mechanism that might have produced them. The models are thus supermodels of many standard ones with mechanistic parameterizations. We study phylogenetic invariants for the models and establish identifiability of species tree topologies using them.

Key words. multispecies coalescent, SVDquartets, flattening matrix, phylogenetic invariants, generic identifiability

MSC codes. 92B10, 92D15, 62R01

DOI. 10.1137/23M1558872

1. Introduction.

The SVDquartets method of Chifman and Kubatko [4, 5] initiated a novel framework for species tree inference from genomic-scale data. Recognizing that individual sites may evolve along different "gene trees" due to the population-genetic effect of incomplete lineage sorting, their method is designed to work with site pattern data generated by the multispecies coalescent model of this process combined with a standard model of nucleotide substitutions across sequence sites. However, rather than try to associate particular gene trees to sites, they regard the observed site pattern distribution as a *coalescent mixture*. This effectively integrates the individual gene trees out of the analysis and allows them to formulate statistical tests based on an algebraic understanding of the site pattern frequencies. These tests detect the unrooted species tree topology in the case of four taxa. For a larger set of taxa, species trees can be found by inferring each quartet and then applying some method of quartet amalgamation. This leads to their SVDquartets method of species tree inference, which is implemented in PAUP* [16] and which continues to be an important tool for practical phylogenetic inference (e.g., [10, 14, 6]).

*Received by the editors March 14, 2023; accepted for publication (in revised form) November 13, 2023; published electronically February 15, 2024.
https://doi.org/10.1137/23M1558872
Funding: The first and third authors were partially supported by NSF grant 2051760 and by NIH grant P20GM103395.
†Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK 99775 USA (e.allman@alaska.edu, j.rhodes@alaska.edu).
‡Department of Mathematical & Computational Sciences, The College of Wooster, Wooster, OH 44691 USA (clong@wooster.edu).

The inference of unrooted 4-taxon species tree topologies in the SVDquartets approach is based on an algebraic insight that a certain flattening matrix built from the site pattern distribution should have low rank on a distribution exactly arising from the model. The mathematical arguments for this in [5] are based on the existence of a rooted cherry (i.e., a 2-clade) on an ultrametric species tree, leading to a symmetry in the site pattern distribution. Since any rooted 4-taxon tree with unrooted topology $ab|cd$ must display at least one of the clades $\{a, b\}$ or $\{c, d\}$, detecting that one or both of these clades is present is equivalent to determining the unrooted tree. The SVDquartets method tests precisely this, without determining which of the clades is present.

In this work, we examine the algebraic framework underlying the work of Chifman and Kubatko and its subsequent extensions. We observe that the symmetry conditions implied by the Chifman–Kubatko model are key to their inference approach. Based on this observation, we formulate several statistical models, encompassing those of [5] as well as several more general mechanistic models, which capture the fundamental assumptions needed to justify SVDquartets. In contrast with the sorts of models generally used in empirical phylogenetics, which have a mechanistic interpretation (e.g., generation of gene trees by the coalescent process, generation of sequences by site-substitution models on the gene trees), the models here have only a descriptive interpretation, as they are defined algebraically by constraints on site pattern distributions.

An important consequence of defining our models in this way is that it becomes more clear that SVDquartets can give consistent species tree inference for mechanistic mixture models much more general than that described in [5], as hinted by results in [11, 12]. In fact, it is easy to formulate plausible mechanistic models with many parameters (e.g., mixtures with many different nucleotide substitution processes) for which many of the numerical parameters must be nonidentifiable, yet which lie within our models. Our work shows that the tree topology is nonetheless identifiable for generic numerical parameters, and in particular that inference of the topology by SVDquartets is statistically consistent. Consistency of tree topology inference by other methods, such as maximum likelihood applied to our new models, should similarly follow, despite nonidentifiability of the mechanistic numerical parameters. Such generality should be viewed as a strength of SVDquartets, as some of the model misspecification arising from the assumption of a simple substitution process across the entire genome is avoided.

A second consequence is that our models highlight a symmetry in the site pattern distribution that reflects the *rooted* species tree, a symmetry that is present even for 3-taxon trees. Methods for inference of the species tree root in the same framework were proposed in [9, 17], but both of these works considered four taxa at a time, which is the smallest *unrooted* tree size in which topologies may differ. Since rooted trees are determined by their rooted triples, focusing on the 3-taxon case offers clear advantages for developing new inference methods. Unfortunately, in doing so, we lose the ability to naturally base statistical inference on rank conditions on matrices of the sort that underlie SVDquartets. Indeed, the possible flattening matrices for DNA site pattern data from the Chifman–Kubatko model in the 3-taxon case are all 4×16 with full rank, so rank alone cannot distinguish them. As a consequence, the matrix singular value decomposition (SVD) of the flattening matrix, which is used to determine approximate rank in the SVDquartets method, has no obvious role. However, we present

an alternative matrix that must satisfy certain rank conditions in the 3-taxon case, which suggests it may be possible to develop a 3-taxon method analagous to SVDquartets.

Our work here is theoretical, dealing primarily with model definitions and algebraic consequences of those models. We suggest its implications for data analysis, but do not explore possible methods based on these results in depth. We begin the next section with a review of the model of [5] and use it to motivate the introduction of our first model, the *ultrametric exchangeable model.* We then discuss a number of its submodels on ultrametric trees, and show in section 3 that the species tree parameter of these models is generically identifiable and species tree inference by SVDquartets is justified for all. In section 4, we give a recursive formula for computing the dimension of the ultrametric exchangeable model, in terms of the dimensions of its subtree models joined at the root. This indicates that the dimension depends on the topology of the tree, which has implications for inference methods. In section 5, we drop the assumption of an ultrametric species tree, reviewing the model of [12] in this setting and using it to motivate our second model, the *extended exchangeable model.* In section 6, we explore the extended exchangeable model in more depth by restricting to 3-taxon trees and determining several algebraic invariants of this model. Finally, in section 7, we show that the species tree parameter of the extended exchangeable model, as well as those of several mechanistic models that it contains, is generically identifiable.

For general references to nonalgebraic aspects of mathematical phylogenetics we recommend [15] for combinatorial material and [8] for background on models.

2. A genomic model of site patterns on ultrametric trees.

We begin by reviewing the simplest mechanistic model of Chifman and Kubatko [5]. For emphasis, we call this model (and others) *mechanistic* since it is formulated through stochastic models of the processes of incomplete lineage sorting and of nucleotide substitution. Many mechanistic models, including that of [5], will be included as submodels of the more general nonmechanistic models we define below and for which the theory underlying SVDquartets applies more broadly.

Specifically, let $\sigma^+ = (\psi^+, \lambda)$ be a rooted species tree on a set of taxa X, with rooted leaf-labeled topology ψ^+ and edge lengths λ in number of generations. The "+" in this notation emphasizes that the tree is rooted, as is needed for the coalescent model. Assume σ^+ is ultrametric, that is, all its leaves are equidistant from the root. Let N be a single constant population size for all populations (i.e., edges) in the species tree, and let μ be a single scalar substitution rate for all populations. For a nucleotide substitution model, fix some general time-reversible (GTR) rate matrix Q with associated stable base distribution π [8].

These parameters determine a DNA site pattern distribution as follows: a site is first assigned a leaf-labeled ultrametric gene tree T sampled under the multispecies coalescent model on σ^+ with population size N, with one gene lineage sampled per taxon. Then a site evolves on T according to the nucleotide substitution model with root distribution π and rate matrix μQ. Site patterns thus have a distribution which is a *coalescent independent mixture* of site pattern distributions arising from the same GTR model on individual gene trees. We denote this model by $\mathrm{CK} = \mathrm{CK}(\sigma^+, N, \mu, Q, \pi)$. (While CK has a mild nonidentifiability issue in that λ, N, and μ are not separately identifiable, this will not be of concern in this work since our focus is on inferring the topology ψ^+.)

A key feature of the CK model is an exchangeability property that it inherits from the multispecies coalescent, due to the nature of the substitution model. Specifically, suppose $\{a,b\} \subseteq X$ is a 2-clade displayed on σ^+. Then for any metric gene tree T, let T' be the gene tree obtained from T by switching the labels a and b. Then the ultrametricity of σ^+ together with exchangeability of lineages under the coalescent model implies T and T' are equiprobable. Now consider any site pattern $z = (z_1, z_2, \ldots, z_n)$ for X, where $z_i \in \{A, G, C, T\}$ is the base for taxon $x_i \in X$, and let z' be the site pattern with the a and b entries of z interchanged. Then under the base substitution model the probability of z on T equals the probability of z' on T'. Thus, with $\mathcal{T}$ denoting the space of all metric gene trees T on X,

$$\begin{aligned}\mathbb{P}(z \mid \sigma^+, N, \mu, Q, \pi) &= \int_{\mathcal{T}} \mathbb{P}(z \mid T, \mu, Q, \pi)\mathbb{P}(T \mid \sigma^+, N)\, dT \\ &= \int_{\mathcal{T}} \mathbb{P}(z' \mid T', \mu, Q, \pi)\mathbb{P}(T' \mid \sigma^+, N)\, dT' \\ &= \mathbb{P}(z' \mid \sigma^+, N, \mu, Q, \pi).\end{aligned} \tag{2.1}$$

Thus any 2-clade on the species tree produces symmetry in the site pattern frequency distribution. Moreover, since both the multispecies coalescent model and the sequence substitution model are well behaved with respect to marginalizing over taxa, it immediately follows that 2-clades on the induced subtrees $\sigma^+|_Y$ on subsets $Y \subset X$ will produce symmetries in the marginalizations of the site pattern distribution to Y.

This motivates the following definition of an algebraic model of site pattern probabilities. In this definition and in what follows, it will be convenient to regard a site pattern probability distribution P from a κ-state model on an n-leaf tree as an *n-way site pattern probability tensor*. That is, we regard $P = (p_{i_1 \ldots i_n})$ as a $\kappa \times \cdots \times \kappa$ array with nonnegative entries adding to 1, where $p_{i_1 \ldots i_n}$ denotes the probability that the n (ordered) taxa are in state $(i_1, \ldots, i_n)$.

Definition 2.1. *Let ψ^+ be a rooted binary topological species tree on X, and $\kappa \geq 2$. Then the κ-state ultrametric exchangeable model, $UE_\kappa(\psi^+)$, is the set of all $\kappa \times \kappa \times \cdots \times \kappa$ $|X|$-way site pattern probability tensors P, such that for every $Y \subseteq X$, and every 2-clade $\{a,b\}$ on $\psi^+|_Y$, the marginal distribution P_Y of site patterns on Y is invariant under exchanging the a and b indices. The collection of all distributions as ψ^+ ranges over rooted binary topological trees on X is the UE model (or the UE_κ model to avoid ambiguity).*

Although this model has "ultrametric" in its name, note that the tree ψ^+ is a topological rooted tree, with no edge lengths. "Ultrametric" here refers to the motivation for the model, generalizing the CK model on an ultrametric species tree discussed above. While one can contrive mechanistic models on nonultrametric trees that lead to distributions in the $\mathrm{UE}_\kappa(\psi^+)$ model, we do not find them very natural and prefer to highlight the ultrametricity that a plausible mechanistic model is likely to require to lie within $\mathrm{UE}_\kappa(\psi^+)$.

It is important to note that unlike most models in phylogenetics, including the CK model above, the UE model is not defined through mechanistically interpretable parameters. Rather it has a descriptive form relating entries of the model's joint distributions, chosen to reflect certain implicit features of the CK model. The UE model then can be viewed as a relaxation, or supermodel, of that more restrictive model.

Example 2.2. Let ψ^+ be the rooted 3-taxon tree $(a,(b,c))$. (See [8] for an explanation of Newick notation for trees.) Consider a 2-state substitution model with states $\{0,1\}$. A probability distribution for the UE(ψ^+) model is $P=(p_{ijk})$, a $2\times 2\times 2$ array with entries the joint probabilities for assignments of states to the taxa, $p_{ijk}=\mathbb{P}(a=\text{`}i\text{'},b=\text{`}j\text{'},c=\text{`}k\text{'})$.

Since the constraints on the model arise only from subsets $Y\subseteq\{a,b,c\}$ that contain at least two taxa, there are four subsets of interest:

$$\{a,b,c\},\ \{b,c\},\ \{a,b\},\ \{a,c\}.$$

Then $\mathrm{UE}_2(\psi^+)$ is a subset of the probability simplex $\Delta^7\subset\mathbb{R}^8$ defined by the following linear equations:

$$\begin{aligned}\{a,b,c\}:&\ \begin{cases}p_{010}=p_{001},\\ p_{101}=p_{110},\end{cases}\\ \{b,c\}:&\ p_{001}+p_{101}=p_{010}+p_{110},\\ \{a,b\}:&\ p_{010}+p_{011}=p_{100}+p_{101},\\ \{a,c\}:&\ p_{001}+p_{011}=p_{100}+p_{110}.\end{aligned}$$

The first two constraints, for $\{a,b,c\}$, express that slices on the first index of probability tensors in $\mathrm{UE}_2(\psi^+)$ are symmetric. Specifically, if the slice $P_{i\cdot\cdot}$ denotes the conditional distribution of b,c when a is in state i, then the 2×2 matrix $P_{i\cdot\cdot}$ is symmetric for each $i\in\{0,1\}$. These imply the third equation, for $\{b,c\}$, expressing that marginalizing over the first index gives a symmetric matrix. The fourth equation, for $\{a,b\}$, is independent of the first three, but with them implies the fifth one, for $\{a,c\}$.

Taking into account the probabilistic requirement that $\sum_{i,j,k\in\{0,1\}}p_{ijk}=1$, we see the model is a restriction of a four-dimensional affine space to the simplex Δ^7 with $0\le p_{ijk}\le 1$.

Note that far more complicated models of site pattern evolution on a species tree than the CK model give rise to distributions within the UE model, since the only requirement is that the resulting site pattern distributions reflect the symmetries of the species tree. For instance, [5] gives an extension allowing for Γ-distributed rate variation across sites. A further generalization, allowing for edge-dependent variation of the population size $N=N_e$, as well as time-dependent variation in the substitution rate μ across the species tree, can easily be seen to produce distributions lying within UE. Since time-reversibility was not used in the above derivation of the symmetry conditions for the CK model, that condition on the substitution rate matrix can be dropped to obtain a more general model that lies within UE. It has also been shown that a model of gene flow on a 3-taxon ultrametric species tree produces site pattern probability distributions that reflect the symmetry in the 2-clade of the species tree [11, Proposition 0.8].

Importantly, the UE(ψ^+) model for a fixed tree ψ^+ is a convex set, since it can be expressed as the solution set of a system of linear equations and inequalities. It immediately follows that very general mixtures of submodels of the UE(ψ^+) model, which are simply convex sums or integrals of distributions in the submodels, are also instances of the UE model on the same tree. This observation immediately recovers that the GTR + Gamma coalescent mixture of [5] lies in UE from our earlier derivation for the CK model, and an analogous result for an

arbitrary rate distribution. Similarly, the mechanistic models in [1] on ultrametric trees that allow for variation in the substitution rate matrix across sites also are submodels of UE. In focusing on the UE model we obtain results that apply to all these models, and possibly more to be formulated in the future.

3. Generic identifiability of trees under the UE model. To use a statistical model for valid inference, it is necessary that any parameter one wishes to infer is *identifiable*; that is, a probability distribution from the model must uniquely determine the parameter. For phylogenetic and other complex models, this strict notion is generally too strong to hold, but one can often establish a similar generic result, that the set of distributions on which identifiability fails is of negligible size (measure zero) within the model. This permits inference after a slight restriction of the parameter space.

Note that the parameter space of the UE model for fixed X and κ is simply the set of probability distributions in the UE model, a union over all ψ^+ of the convex subsets $UE(\psi^+)$ of Euclidean space. It therefore has a natural measure and notion of genericity.

Theorem 3.1. *The rooted binary topological tree ψ^+ is identifiable from a generic probability distribution in the UE model.*

Proof. Fix κ and a taxon set X. Since for each binary species tree topology ψ^+ the symmetry conditions are expressible by linear equations, the UE model for ψ^+ is the intersection of a linear space with the probability simplex. We establish the result by showing that the linear model spaces for different ψ^+ are not contained in one another, since then their intersection is of lower dimension and hence of measure zero within them.

That the linear spaces are not contained in one another will follow by establishing that for each ψ^+ there is at least one distribution in $\mathrm{UE}_\kappa(\psi^+)$ that fails to have any "extra symmetry" required for it to be in the model for a different tree. To construct such a distribution, assign positive edge lengths to ψ^+ so that the tree is ultrametric, and consider on it a κ-state analogue of the (noncoalescent) Jukes–Cantor [8] model (henceforth denoting this analogue "JC") in which the base distribution is uniform and all off-diagonal entries of the rate matrix are equal and positive. The resulting site pattern distribution P is easily seen to have the necessary symmetries to lie in the UE model: For a, b forming a 2-clade on $\psi^+|_Y$ the conditional distribution fixing states at other taxa on $\psi^+|_Y$ has the form $M^T \mathrm{diag}(w) M$. Here M is a JC Markov matrix describing substitutions on each of the pendant edges to a, b and w is a vector giving the distribution of states at the most recent common ancestor of a, b conditioned on the states at other taxa.

To show P has no extra symmetries, suppose to the contrary that there is a $Y \subset X$ containing two taxa a, c where $P|_Y$ is invariant under exchanging the a and c indices, yet a, c do not form a cherry on $\psi^+|_Y$. Then, after possibly interchanging the names of a, c, there is a third taxon b such that the rooted triple $((a, b), c)$ is displayed on $\psi^+|_Y$. Moreover, by further marginalizing to $Y' = \{a, b, c\}$, we have that $P|_{Y'}$ arises from a JC model on a 3-taxon ultrametric tree with positive edge lengths and rooted topology $((a, b), c)$ and exhibits a, c symmetry.

To see that this is impossible, note that if $P|_{Y'}$ has both a, b and a, c symmetry, then it also exhibits b, c symmetry. Thus, all marginalizations of $P|_{Y'}$ to two taxa are equal. This implies all JC distances between taxa, which can be computed from these marginalizations, are equal. This contradicts that the tree was binary. ∎

Note that the proof above did not consider a coalescent process in any way in order to show that extra symmetries do not generically hold in UE (ψ^+). However, since applications may consider submodels of the UE model, such as the CK model, it is necessary to ensure they do not lie within the exceptional set of nongeneric points in the UE model where tree identifiability may fail. To address this issue, we seek an identifiability result for more general mechanistic models that have an *analytic parameterization,* by which we mean that for each topology ψ^+ there is an analytic map from a full-dimensional connected subset of $\mathbb{R}^k$, for some k, to the set of probability distributions comprising the model. For example, if σ^+ is a rooted metric species tree with shape ψ^+, and site pattern frequency distributions are generated on gene trees arising under the coalescent using the GTR + I+Γ model, then the collection of such distributions is given by an analytic parameterization and as such is a submodel of UE(ψ^+).

Since the UE model is viewed as a subset of Euclidean space, it inherits the standard topology, which allows us to consider limit points of submodels.

Theorem 3.2. *Consider any submodel of the UE model with an analytic parameterization, which is general enough to have the distributions of the JC model as limit points. Then for generic parameters of that submodel the rooted topological tree ψ^+ is identifiable.*

Proof. Let

$$f : \Theta \to UE(\psi^+)$$

denote the parameterization map for the submodel on tree ψ^+. Then $f(\Theta)$ cannot lie entirely in $UE(\phi^+)$ for any $\phi^+ \neq \psi^+$, since, as shown in the previous proof, there are points from the JC model in the closure of $f(\Theta)$ which are not in the closed set $UE(\phi^+)$. Thus the set $f^{-1}(UE(\psi^+) \cap UE(\phi^+))$ is a proper analytic subvariety of Θ, and hence of measure zero in it. Since there are only finitely many ϕ^+, for generic points in Θ the resulting distribution lies in the UE model for ψ^+ only. ∎

Note that the CK model, which is analytically parameterized, has the JC model as limit points, since after choosing a JC substitution process one can let the population size $N \to 0^+$. This effectively "turns off" the coalescent process, as small population sizes result in rapid coalescence.

As noted earlier, the UE model includes distributions arising from many standard mechanistic models on ultrametric trees, as well as mixtures of them with many more components than are typically considered in practical inference due to computational demands. Many of these submodels (e.g., a mixture of a large number of GTR + Gamma coalescent mixtures with different rate matrices, or a similar model dropping the requirement of time-reversibility) contain the JC model. Thus the above theorem applies to show the rooted topological tree is generically identifiable. Even for mixtures which have so many mechanistic numerical parameters that dimension arguments show they cannot all be individually identifiable, the species tree topology remains so. This is a potentially valuable observation, as a scheme designed for inference of a tree under the UE model may avoid some issues of model misspecification that might arise with a more standard approach of restricting to a very simple model so that all mechanistic numerical parameters are identifiable as well.

The above theorems of course imply the weaker statement that for the UE model (and many analytic submodels of the UE model) on four or more taxa, the unrooted species tree topology is generically identifiable. We next turn to showing that the statistical consistency of SVDquartets for inferring the unrooted tree under the CK model extends to consistency under the UE model. The first step is to prove that the specific flattening matrices considered in the SVDquartets method satisfy certain rank conditions, the content of the next theorem.

Recall that if a $\kappa \times \kappa \times \kappa \times \kappa$ array P has indices corresponding to taxa a, b, c, d, then the flattening $\text{Flat}_{ab|cd}(P)$ is a $\kappa^2 \times \kappa^2$ matrix with row and column indices in $\kappa \times \kappa$ and $((i,j),(k,l))$-entry p_{ijkl}.

Theorem 3.3. *For $P \in UE_\kappa(\psi^+)$, and $ab|cd$ any unrooted quartet induced from the tree ψ^+, let $\tilde{P} = P|_{\{a,b,c,d\}}$ denote the marginalization to the taxa a, b, c, d. Then for all such P, $\text{Flat}_{ab|cd}(\tilde{P})$ has rank at most $\binom{\kappa+1}{2}$, while for generic P, $\text{Flat}_{ac|bd}(\tilde{P})$ and $\text{Flat}_{ad|bc}(\tilde{P})$ have rank κ^2.*

Proof. Since $\psi^+|_{\{a,b,c,d\}}$ has at least one cherry, assume one is formed by a, b. Then symmetry under exchanging the a, b indices of $\tilde{P}$ shows that for each $1 \le i < j \le \kappa$, the (i,j) and (j,i) rows of $\text{Flat}_{ab|cd}(\tilde{P})$ are identical. Thus that flattening has at most $\kappa^2 - \binom{\kappa}{2} = \binom{\kappa+1}{2}$ distinct rows, and its rank is at most $\binom{\kappa+1}{2}$.

We prove the second statement for $\text{Flat}_{ac|bd}(\tilde{P})$, noting that the argument for $\text{Flat}_{ad|bc}(\tilde{P})$ is similar. To show that for generic $P \in \text{UE}_\kappa(\psi^+)$, $\text{Flat}_{ac|bd}(\tilde{P})$ has full rank, it suffices to construct a single P for which this flattening matrix is full rank. To see that this is the case, consider the algebraic variety

$$\mathcal{V}_{ac|bd} = \{P \in \mathbb{R}^{\kappa^{|X|}} \,|\, \det(\text{Flat}_{ac|bd}(\tilde{P})) = 0\}.$$

This variety is defined by a single degree κ^2 polynomial and contains all of the points P for which $\text{Flat}_{ac|bd}(\tilde{P})$ is singular. If there is a single point $P \in UE_\kappa(\psi^+)$ for which $\det(\text{Flat}_{ac|bd}(\tilde{P})) \neq 0$, then the affine space $UE_\kappa(\psi^+)$ is not contained in $\mathcal{V}_{ac|bd}$. Thus, the intersection of $UE_\kappa(\psi^+)$ with $\mathcal{V}_{ac|bd}$ is a proper subvariety of $UE_\kappa(\psi^+)$, and hence of measure zero within it. Thus, generically, $\text{Flat}_{ac|bd}(\tilde{P})$ is full rank.

To construct such a probability distribution, assign any positive lengths to the edges of ψ^+ so that it becomes ultrametric, and consider the κ-state JC model on it (with no coalescent process). This leads to a distribution $P \in \text{UE}_\kappa(\psi^+)$. Then $\tilde{P}$ arises from the Jukes–Cantor model on the induced rooted 4-taxon tree. Since the JC model is time reversible, $\tilde{P}$ is also obtained by rooting the quartet tree at the most recent common ancestor of a and b, with nonidentity JC Markov matrices on each of the five edges of this rerooted tree. Let M_a, M_b, M_c, M_d denote the Markov matrices on the pendant edges and M_{int} on the internal edge, so that $F = (1/\kappa)M_{int}$ is the distribution of pairs of bases at the endpoints of the internal edge. Let $N_{ac} = M_a \otimes M_c$ and $N_{bd} = M_b \otimes M_d$ denote the Kronkecker products. Then, following the details of [2, section 4], the flattening matrix may be expressed as

$$\text{Flat}_{ac|bd}(\tilde{P}) = N_{ac}^T D N_{bd},$$

where D is a $\kappa^2 \times \kappa^2$ diagonal matrix formed from the entries of F.

Since M_{int} is assumed to be a nonidentity JC matrix, F has no zero entries, so D has rank κ^2. Similarly, the JC transition matrices M_a, M_b, M_c, M_d are nonsingular, and since the Kronecker product of nonsingular matrices is nonsingular, so are N_{ac}^T and N_{bd}. Thus $\text{Flat}_{ac|bd}$ generically has full rank. ∎

The argument in this proof, that generically the ranks of "wrong" flattenings of quartet distributions are large, proceeded by constructing an element of the UE model using a parameterized model in the absence of a coalescent process. However, just as was done in Theorem 3.2, we can extend the conclusion to analytic submodels of the UE model, such as those incorporating the coalescent. For instance, since the CK model has the noncoalescent JC model as a limit, this implies that there are points in the CK model that are arbitrarily close to the point P constructed in the proof, which therefore must also have rank κ^2 flattenings, as matrix rank is lower semicontinuous. We can thus obtain the following generalization of a result from [5].

Theorem 3.4. *Consider any submodel of the $UE(\psi^+)$ model with an analytic parameterization general enough to have the distributions of the JC model as limit points. If ψ^+ displays the quartet $ab|cd$, then for all distributions P in the model, with $\tilde{P} = P|_{\{a,b,c,d\}}$, $\text{Flat}_{ab|cd}(\tilde{P})$ has rank at most $\binom{\kappa+1}{2}$, while for generic P, $\text{Flat}_{ac|bd}(\tilde{P})$ and $\text{Flat}_{ad|bc}(\tilde{P})$ have rank κ^2.*

We note that our proof of this theorem has avoided the explicit calculations and more intricate arguments that appear in [5] while also establishing the result in a more general setting. This is possible because of our use of a tensor P in the closure of the CK model, but not in the CK model, as well as adopting the viewpoint of [2] on flattenings as matrix products.

Using the two preceding theorems on identifiability, the statistical consistency of the SVDquartets method can be obtained. When Chifman and Kubatko [5] proved essentially the same result on ranks of flattenings for the CK model, they highlighted it as an identifiability result, but did not explicitly make a claim of consistency. The consistency result for SVDquartets was then unambiguously stated and proved in this setting in [18], which also gave an analysis of the convergence rate.

Here we show that their argument for the consistency of SVDquartets applies more generally to site patterns generated under the UE model, as well as many submodels. In particular, it validates the consistency of inference under models allowing mixtures of coalescent mixtures which may have different substitution processes across the genome, as described in [1].

To be precise, we must first specify a method of quartet amalgamation $\mathcal{M}$, which takes a collection of quartet trees (one for each 4-taxon subset of X) and produces an unrooted topological tree on X. In order to establish consistency, we require that if all quartet trees in the collection given to the method $\mathcal{M}$ are displayed on a common tree T on X, then $\mathcal{M}$ returns T. Following [13], we say such a method is *exact* while recognizing that for large sets X one generally must use a heuristic method $\mathcal{M}'$ that seeks to approximate $\mathcal{M}$.

Theorem 3.5. *The SVDquartets method, using an exact method to construct a tree from a collection of quartets, gives a statistically consistent unrooted species tree topology estimator for generic parameters under the UE model, and under any submodel with an analytic parameterization general enough to have the JC model as a limit.*

Proof. To simplify notation in the argument, let $\text{Flat}_{ac|bd}(P)$ denote the $ac|bd$ flattening of the marginalization $P|_{\{a,b,c,d\}}$.

By Theorems 3.3 and 3.4 for generic parameters giving a probability distribution P in the model and any four taxa a, b, c, d such that $ab|cd$ is displayed on the unrooted tree ψ, $\text{Flat}_{ab|cd}(P)$ has rank at most $\binom{\kappa+1}{2}$, while $\text{Flat}_{ac|bd}(P)$ and $\text{Flat}_{ad|bc}(P)$ have rank κ^2. This implies that $\text{Flat}_{ab|cd}(P)$ will have at least $\binom{\kappa}{2}$ singular values of 0, while $\text{Flat}_{ac|bd}(P)$ and $\text{Flat}_{ad|bc}(P)$ have all positive singular values. For a finite sample of s sites from the model, denote the empirical distribution by $\hat{P}_s$. Then for any $\epsilon > 0$ and any norm $|\cdot|$,

$$\lim_{s\to\infty} \Pr\left(|\hat{P}_s - P| < \epsilon\right) = 1.$$

Since the vector $\sigma(M)$ of ordered singular values of a matrix M is a continuous function of the matrix, this implies that for each $q \in \{ab|cd, ac|bd, ad|bc\}$

$$\lim_{s\to\infty} \Pr\left(\|\sigma(\text{Flat}_q(\hat{P}_s)) - \sigma(\text{Flat}_q(P))\| < \epsilon\right) = 1,$$

where $\|\cdot\|$ denotes any vector norm. With the SVD score $\mu(M)$ defined as the sum of the $\binom{\kappa}{2}$ smallest singular values of a $\kappa^2 \times \kappa^2$ matrix M, we know

$$0 = \mu(\text{Flat}_{ab|cd}(P)) < \min\left\{\mu(\text{Flat}_{ac|bd}(P)),\ \mu(\text{Flat}_{ad|bc}(P))\right\}.$$

But it then follows that

$$\lim_{s\to\infty} \Pr\left(\mu(\text{Flat}_{ab|cd}(\hat{P}_s)) < \min\left\{\mu(\text{Flat}_{ac|bd}(\hat{P}_s)), \mu(\text{Flat}_{ad|bc}(\hat{P}_s))\right\}\right) = 1.$$

Thus, as the sample size s grows, the probability that choosing the quartet tree on a, b, c, d minimizing μ gives the quartet tree displayed on ψ approaches 1.

Since this probability approaches 1 for each set of four taxa, and there are only finitely many such sets, the probability that all quartet trees inferred by minimizing μ are displayed on the species tree approaches 1. Thus with probability approaching 1, the method M will return the correct species tree. ∎

Remark 3.6. Since the model $\text{UE}(\psi^+)$ is defined by linear equalites and inequalities, it would be natural to use that structure in inference. However, the model of interest for tree inference is

$$UE = \cup_{\psi^+} UE(\psi^+),$$

where the union is taken over all rooted topological trees ψ^+ that might relate the taxa. Finding the maximum likelihood estimate of the tree involves maximizing the likelihood of the data for every such ψ^+, and then choosing the tree with the highest maximum. Thus the well-understood issue in phylogenetics of the large size of tree space when many taxa are under consideration means many individual linear models would need to be considered.

SVDquartets, by first inferring quartet trees, handles the many-taxon tree problem by seeking a large tree that is most consistent with inferred quartets. We note that a third

approach is also possible. For each set of three taxa, one of the three possible rooted topological trees which might relate them could be inferred by likelihood, and then some rooted triple method used to choose a large tree.

Which of these approaches might be preferable, in terms of both statistical accuracy and computational performance, we do not address.

4. Dimension of UE models on large trees. Although the symmetry conditions of the UE model have been expressed as linear constraint equations, these constraints are not in general independent, as was shown for a particular 3-taxon species tree in Example 2.2. In that example, it was easy to determine a basis of constraints, and thus the dimension of the model. In this section we investigate larger trees and determine the model dimension.

Knowledge of dimension is important for several reasons. First, it gives us a basic insight into how restrictive the model on a particular tree topology is. Second, if one is to use these models for tree inference, the dimension is important for judging how close a data point is to fitting the model. Intuitively, data is conceptualized as coming from a true model point with "noise" added, and if a model has high dimension the noise tends to do less to move that data from the model than if it had lower dimension. Such dimensionality considerations are made rigorous in many model selection criteria, for instance, the Akaike information criterion and the Bayesian information criterion.

For a rooted topological tree ψ^+ on taxa X we consider the model $\mathrm{UE}_\kappa(\psi^+)$. Let $d_\kappa(\psi^+)$ denote the dimension of the affine space $\mathcal{V}(\psi^+) \subset \mathbb{R}^{\kappa^{|X|}}$ of all tensors satisfying the linear equations expressing the symmetry conditions defining the model, as well as that all entries of the distribution tensor sum to 1, but with no nonnegativity requirements (i.e., the affine, or Zariski, closure of the model). By dropping the condition that tensor entries sum to 1, we pass to the cone, a linear space $L(\psi^+)$ of dimension $c_\kappa(\psi^+) = d_\kappa(\psi^+) + 1$. We now give a recursive formula for computing the dimension $c_\kappa(\psi^+)$.

Theorem 4.1. *For a rooted binary topological tree ψ^+ on a taxon set X, let ψ^+_A and ψ^+_B be the rooted subtrees descendant from the child nodes of the root of ψ^+, on taxa A and B, respectively, so that $X = A \sqcup B$ and in Newick notation $\psi^+ = (\psi^+_A, \psi^+_B)$. Then*

$$c_\kappa(\psi^+) = c_\kappa(\psi^+_A)c_\kappa(\psi^+_B) - \binom{\kappa}{2}.$$

The proof will follow after preliminary terminology, comments, and a lemma. For any multidimensional tensor or array A, a subarray determined by fixing a subset of the indices is referred to as a *slice* of A, so that fixing all but two indices determines a *matrix slice.* Given a node v on a rooted, and hence directed, tree with leaves labeled by a set of taxa X, the subset of X of *descendants of* v is denoted $\mathrm{desc}_X(v)$.

For a topological rooted species tree ψ^+ on X, we can construct a set of equations defining the cone $L(\psi^+)$ by considering every subset $Y \subseteq X$ and every 2-clade $\{a, b\}$ of each $\psi^+_{|Y}$ as was done in Example 2.2. However, as we saw in that example, the equations we obtain in this way are not necessarily independent. As a first step toward proving Theorem 4.1, we construct a smaller (though still not necessarily independent) set of linear equations defining the cone $L(\psi^+)$. This set is defined by associating a set of linear equations to each vertex of the topological rooted tree ψ^+ on X. Specifically, for each internal vertex v of ψ^+ choose two

taxa a, b which have v as their most recent common ancestor. Let P be an $|X|$-dimensional $\kappa \times \cdots \times \kappa$ tensor of indeterminates, with indices corresponding to taxa in X, and let P_{ab} denote the marginalization of P over all indices corresponding to taxa in $\mathrm{desc}_X(v) \setminus \{a, b\}$. Each choice of the indices corresponding to taxa in $X \setminus \mathrm{desc}_X(v)$ determines a matrix slice of P_{ab}, with indices corresponding to a, b. Expressing that each of these slices is a symmetric matrix yields a collection of linear equations. Denote this set of equations by $\mathcal{S}_v = \mathcal{S}(\psi^+, \{a, b\})$. Though the set $\mathcal{S}_v$ will depend on the particular pair of taxa (a, b) chosen, for our purposes the particular pair is irrelevant, so one can designate any consistent rule for selecting the pair (a, b) so that the $\mathcal{S}_v$ are well-defined. If v is not an internal vertex of ψ^+, define S_v to be the empty set.

How this construction applies to Example 2.2 is discussed after its utility is shown in the following lemma.

Lemma 4.2. *Let ψ^+ be a topological rooted tree on X. Then the set*

$$\mathcal{S} = \bigcup_{v \in \mathcal{V}(\psi^+)} \mathcal{S}_v$$

defines the cone $L(\psi^+)$.

Proof. It is enough to show that if v is the most recent common ancestor (MRCA) of taxa a, b as well as of taxa a, c, then the linear equations expressing symmetry of slices of P_{ac} are contained in the span of those expressing symmetry of slices of P_{ab} together with those equations in $\mathcal{S}$ arising from nodes descended from v. We show this inductively, proceeding from the leaves of the tree to the root. The base case, when v has only two leaf descendants, is trivial. Assume the result holds for the internal nodes descended from v. Let the children of v be v_1, which is ancestral to or equal to a, and v_2, which is ancestral to b, c since ψ^+ is binary. Then $w = \mathrm{MRCA}(b, c)$ is a descendent of v_2. The equations arising from w express that any entry of the marginalization of P over all descendants of w except b, c is invariant under exchanging the b, c indices. Since the entries of P_{ab} arise from further marginalization, the equations expressing symmetry of the ab-slices together with those arising from w imply those expressing the ac-slices of P_{ac} are symmetric. ∎

This proof explains the dependence of the equations in Example 2.2. The $\{a, b, c\}$ constraints are the equations arising from $\mathrm{MRCA}(b, c)$, which in that example required no marginalization of P. The $\{a, b\}$ constraints are the equations arising from the root of the tree that express symmetry of P_{ab} which are obtained by marginalizing P over c. Together, these constraints imply the $\{b, c\}$ and $\{a, c\}$ constraints, the latter of which express symmetry in the slices of P_{ac}.

Proof of Theorem 4.1. Let $n_A = |A|$ and $n_B = |B|$. With $\mathcal{U} = \mathbb{R}^{\kappa^{n_A}}$ and $\mathcal{V} = \mathbb{R}^{\kappa^{n_B}}$, we identify $W = \mathcal{U} \otimes \mathcal{V} = \mathbb{R}^{|X|}$ with the space of $k^{n_A} \times k^{n_B}$ real matrices. In particular, we have $L(\psi_A^+) \subset \mathcal{U}$, $L(\psi_B^+) \subset \mathcal{V}$, and $L(\psi^+) \subset W$.

We first claim that $L(\psi_A^+) \otimes L(\psi_B^+)$ is the subspace $Z \subset W$ defined by the subset $\mathcal{S}'$ of $\mathcal{S} = \mathcal{S}(\psi^+)$ of Lemma 4.2 containing only those equations arising from nonroot internal nodes of ψ^+.

To see $L(\psi_A^+) \otimes L(\psi_B^+) \subseteq Z$, consider an equation in $\mathcal{S}'$ associated to a nonroot node v and its descendant taxa a, b as in the lemma. Without loss of generality, we may assume v is a

node of ψ_A. Then, ordering the taxa so that a, b are the first two, this equation in $\mathcal{S}'$ has the form

$$\sum_{\alpha_1} x_{i,j,\alpha_1,\alpha_2,\beta} - \sum_{\alpha_1} x_{j,i,\alpha_1,\alpha_2,\beta} = 0, \tag{4.1}$$

where the summation over $\alpha_1 \in [k]^m$ runs through all assignments of states to taxa descended from v other than a, b, $\alpha_2 \in [k]^{n_A-2-m}$ is a fixed choice of states for taxa in A not descended from v, $\beta \in [k]^{n_B}$ is a fixed choice of states for the taxa in B, and $i \neq j$. The first subscript $i, j, \alpha_1, \alpha_2, \beta$ is formed by concatenation of these tuples and is thus an element of $[k]^{n_A+n_B}$, as is the second. This equation expresses that column β of a matrix in W satisfies an equation associated to v, a, and b in the definition of $L(\psi_A^+)$. Thus it holds on all of $L(\psi_A^+) \otimes L(\psi_B^+)$, and we obtain the desired inclusion.

To see $L(\psi_A^+) \otimes L(\psi_B^+) \supseteq Z$, note that (4.1) has shown that every column of $z \in Z$ lies in $L(\psi_A^+)$, and likewise every row of z lies in $L(\psi_B^+)$. But from the SVD of z,

$$z = \sum_i c_i \otimes r_i,$$

where the c_i form a basis for the column space of z and the r_i form a basis for the row space of z. Since $c_i \in L(\psi_A^+)$ and $r_i \in L(\psi_B^+)$, it follows that $z \in L(\psi_A^+) \otimes L(\psi_B^+)$, establishing the stated inclusion and that $Z = L(\psi_A^+) \otimes L(\psi_B^+)$.

Now the space $L(\psi^+)$ is the subset of $Z = L(\psi_A^+) \otimes L(\psi_B^+)$ defined by the equations in $\mathcal{S} \setminus \mathcal{S}'$, associated to the root of ψ. To conclude that

$$c_\kappa(\psi^+) = c_\kappa(\psi_A^+) c_\kappa(\psi_B^+) - \binom{\kappa}{2},$$

it is enough to show that we can obtain an independent set of equations defining $L(\psi^+)$ by taking an independent set defining Z and augmenting it by $\binom{\kappa}{2}$ additional independent equations associated to the root.

Let $\mathcal{L}$ be any independent subset of equations in $\mathcal{S}'$ that define Z, and $\mathcal{M} = \mathcal{S} \setminus \mathcal{S}'$ the set of $\binom{\kappa}{2}$ equations associated to the root of ψ^+ (and the choice of $a \in A$ and $b \in B$). Then $\mathcal{L} \cup \mathcal{M}$ defines $L(\psi^+)$. To see that $\mathcal{L} \cup \mathcal{M}$ is independent, first order indices so that a and b indices are listed first among A and B. Then, using "+" in an index to denote the sum over the assignment of all states $[\kappa] = \{1, 2, \ldots, \kappa\}$ in that index, for any $1 \leq i < j \leq k$,

$$x_{i+\cdots+,\, j+\cdots+} - x_{j+\cdots+,\, i+\cdots+} = 0$$

must be the unique element of $\mathcal{M}$ that involves the variable $x_{ii\cdots i,\, jj\cdots j}$ (noting that each equation in $\mathcal{L}$ involves variables that have at least two distinct entries in the indices for A or two distinct entries in the indices for B). Since $\mathcal{L}$ is an independent set, this implies $\mathcal{L} \cup \mathcal{M}$ is independent. ∎

The theorem gives insight into model dimensions for families of "extreme" topologies: rooted caterpillars and fully balanced shapes [15].

Corollary 4.3. *Suppose ψ^+ is a rooted caterpillar tree on n taxa. Then the dimension of the $UE_\kappa(\psi^+)$ model is*

$$d_\kappa(\psi^+) = \frac{\kappa^n + \kappa}{2} - 1.$$

Proof. If $n = 1$, then the model is simply a base distribution for the sole taxon, so $d_\kappa(\psi^+) = \kappa - 1$, consistent with the stated formula. Now inductively assume the stated formula for the rooted caterpillar on $n-1$ taxa. Then by Theorem 4.1, for n taxa

$$c_\kappa(\psi^+) = \left(\frac{\kappa^{n-1} + \kappa}{2}\right)\kappa - \binom{\kappa}{2} = \left(\frac{\kappa^n + \kappa^2}{2}\right) - \left(\frac{\kappa^2 - \kappa}{2}\right) = \frac{\kappa^n + \kappa}{2},$$

and the claim follows. ∎

Also from Theorem 4.1 we can compute that the dimension of the UE model on the 4-taxon balanced tree $((a,b),(c,d))$ is

$$d_\kappa = \left(\frac{\kappa^2 + \kappa}{2}\right)^2 - \binom{\kappa}{2} - 1 = \frac{\kappa(\kappa^3 + 2\kappa^2 - \kappa + 2)}{4} - 1.$$

By comparing the dimensions for the 4-taxon caterpillar and balanced trees, we see that d_k depends on the rooted tree topology, and not only on the number of taxa.

More generally, for a fully balanced tree ψ^+ on $n = 2^\ell$ taxa, Theorem 4.1 yields that

$$d_\kappa(\psi^+) = \mathcal{O}\left(\left(\frac{\kappa(\kappa+1)}{2}\right)^{n/2}\right).$$

Thus for fully balanced trees the dimension is $o(\kappa^n/2)$, while for rooted caterpillars on n taxa, Corollary 4.3 shows the dimension is asymptotic to $\kappa^n/2$. For a fixed number of taxa $n = 2^\ell$, it follows that the dimension of the balanced tree model will be smaller than that of the caterpillar.

This comparison of model dimension for caterpillars and balanced trees is intuitively plausible, as cherries on the full tree lead to more symmetry requirements on a tensor than do cherries on subtrees. In general, the more balanced a tree is, the smaller one might expect the model dimension to be. This leads us to pose the following conjectures, where $RB(n)$ denotes the set of rooted binary n-leaf trees.

Conjecture 4.4. *For all κ, there exists an m such that for $n \geq m$, $d_\kappa(\psi^+)$ is maximized over $\psi^+ \in RB(n)$ when ψ^+ is the n-leaf caterpillar tree.*

Conjecture 4.5. *For all κ, there exists an m such that for $\ell \geq m$, $d_\kappa(\psi^+)$ is minimized over $\psi^+ \in RB(2^\ell)$ when ψ^+ is the 2^ℓ-leaf balanced tree.*

5. A genomic model of site patterns on general trees. In this section, we examine a generalization of the CK model to nonultrametric trees to motivate an algebraic model that encompasses it. Marginalizations (resp., slices) of a site pattern probability tensor will be denoted by placing a "+" (resp., k) in the index summed over (resp., conditioned on). The transpose operator will be denoted with an exponent "T." For example, we can generalize the

equations derived in Example 2.2 for the UE model on the 3-leaf rooted tree $\psi^+ = (a,(b,c))$ for any value of κ using this notation as follows:

(1) $P_{k\cdot\cdot} = P_{k\cdot\cdot}^T$ for $1 \le k \le \kappa$,
(2) $P_{+\cdot\cdot} = P_{+\cdot\cdot}^T$,
(3) $P_{\cdot\cdot+} = P_{\cdot\cdot+}^T$,
(4) $P_{\cdot+\cdot} = P_{\cdot+\cdot}^T$.

For $UE(\psi^+)$, these constraints arise from the taxon subsets (1) $\{a,b,c\}$, (2) $\{b,c\}$, (3) $\{a,b\}$, and (4) $\{a,c\}$, and it is not hard to see that the equations in (1) and (3) imply those in (2) and (4), just as in Example 2.2.

5.1. The extended exchangeability model. In [12], the CK model is extended to permit nonultrametricity of the species tree. This extension allows, for instance, the modeling of relationships between species when generation times or scalar substitution rates differ across populations in the tree. In this same work, flattening matrices are used to establish the generic identifiability of the unrooted species tree topology of the extended model from which it follows that SVDquartets is still a statistically consistent method of inference of the unrooted species tree topology for these models when combined with any exact method of quartet amalgamation.

In order to motivate our algebraic model, first consider a model obtained from the CK by dropping the ultrametricity requirement on the species tree σ^+. Suppose a and b are taxa in a 2-clade on σ^+, and let v be their common parental node. In the special case that the edge lengths of $e_a = (v,a)$ and $e_b = (v,b)$ equal, then the lineages a and b would be exchangeable under this site pattern model as shown for the CK model. Thus, for this particular tree the site pattern distribution can be viewed as a tensor with symmetry in the a and b coordinates. On a general species tree, however, where e_a and e_b may have different lengths and substitution rates may not be consistent, all sites evolve over those edges according to the transition matrices

$$M_a = \exp(s_a Q), \quad s_a = \int_0^{\ell(e_a)} \mu_{e_a}(t)dt,$$
$$M_b = \exp(s_b Q), \quad s_b = \int_0^{\ell(e_b)} \mu_{e_b}(t)dt,$$

where $\ell(e)$ is the length of edge e and $\mu_{e_a}(t)$ and $\mu_{e_b}(t)$ are time-dependent substitution rates.

Supposing, without loss of generality, that $s_a \le s_b$, define the Markov matrix

$$M = M_b M_a^{-1} = \exp((s_b - s_a)Q).$$

Then the site pattern distribution can be viewed as one obtained from a tensor with symmetry in a and b that has been acted on by M in the b-index. More specifically, we imagine that on the edges leading toward both a and b, the Markov matrix M_a describes an initial substitution process, but on the edge to b there is a subsequent substitution process described by M. If we introduce an additional action by M on the edge to a, then in the resulting distribution we would regain symmetry in a and b. Since no coalescent events occur in these pendant edges, there are no complications arising from the coalescent events that do occur.

To formalize this mathematically, suppose P is an N-way $\kappa \times \kappa \times \cdots \times \kappa$ tensor. Define the action of a $\kappa \times \kappa$ matrix M in the kth index of P by $R = P *_k M$ where

$$R(i_1, i_2, \ldots, i_k, \ldots, i_N) = wM,$$

with w the row vector determined by fixing the ℓth index of P to be i_ℓ for all $\ell \neq k$. For example, for $n=3$ and $k=1$, the tensor $P *_1 M$ is specified by $(P *_1 M)_{ijk} = (P_{jk}M)_i$. Given an n-tuple of matrices $(M_1, M_2, \dots M_n)$, let

$$P * (M_1, M_2, \dots, M_n) = (\dots((P *_1 M_1) *_2 M_2) \cdots *_n M_n)$$

denote the action in each of the indices of P.

Definition 5.1. *Let ψ^+ be a rooted topological species tree on X with $|X| = n$. Then the extended exchangeable model, $EE_\kappa(\psi^+)$, is the set of all n-way site pattern probability tensors P such that there is an n-tuple $M = (M_1, M_2, \dots, M_n)$ of $\kappa \times \kappa$ nonsingular Markov matrices M_i and a nonnegative array $\tilde{P}$ in the model $UE_\kappa(\psi^+)$ such that $P * M = \tilde{P}$.*

We note that UE is a submodel of EE: any distribution in $\mathrm{UE}_\kappa(\psi^+)$ is seen to lie in $\mathrm{EE}_\kappa(\psi^+)$ by taking all matrices M_i to be the identity. Also, to ensure that the EE model does not include all distributions, it is important that the M_i be nonsingular in this definition. Otherwise, if the M_i describe processes where *all* states transition to the same state with probability 1, then for any tensor P, $P * (M_1, M_2, \dots M_n) = \tilde{P}$, a tensor with a single diagonal entry equal to 1 that is in UE.

While the UE model on a 2-leaf tree imposes constraints on the probability distribution of site patterns, the 2-leaf EE model is dense among all probability distributions. Indeed, the EE model on such a tree simply requires that the site pattern distribution have the form of $P = M_1^{-T} S M_2^{-1}$ with S a symmetric probability matrix and the M_i Markov. But a dense subset of all probability distributions can be expressed as $P = DM$ for a diagonal matrix D with entries from the row sums of P and an invertible Markov matrix M. We can thus take $M_1 = M$, $S = M^T DM$, and $M_2 = I$.

For a 3-taxon tree, though, the EE model is typically not the full probability simplex Δ^{κ^3-1}. For $\kappa \geq 4$, this follows from a simple dimension bound. The $UE(\psi^+)$ model for a 3-taxon tree ψ^+ has, from Corollary 4.3, dimension

$$d_\kappa = \frac{\kappa^3 + \kappa}{2} - 1.$$

Moreover, the affine closure of the UE model on a 3-taxon tree is mapped to itself by the $*$ action of (M^{-1}, M^{-1}, M^{-1}) for any Markov matrix M. Thus, since nonidentity Markov matrices are needed to act on at most two of the three leaves of the tree, the dimension of the $EE(\psi^+)$ model can be at most

$$\dim(UE(\psi^+)) + 2\kappa(\kappa - 1),$$

where the second term is the number of parameters specifying those matrices. Thus

$$\dim(EE(\psi^+)) \leq \frac{\kappa^3 + 4\kappa^2 - 3\kappa}{2} - 1 < \kappa^3 - 1$$

for all $\kappa \geq 4$.

As we address in Remark 6.4 following Corollary 6.3, we can confirm computationally that for a 3-taxon tree and $\kappa = 3$, $EE(\psi^+)$ is of lower dimension than the probability simplex Δ^{26}, and that for $\kappa = 2$, the Zariski closure of $EE(\psi^+)$ is equal to Δ^7.

Remark 5.2. A more restrictive variant of the EE model, that is closer to the mechanistic models of [12] model, could be defined by requiring that all the "extension" matrices M_i arise as exponentials of the same GTR rate matrix. While this common exponential condition is not expressible purely through algebra, there are other algebraic relaxations of it that one could impose instead, such as that the extension matrices M_i commute and the matrices $\mathrm{diag}(\pi)M_i$ are symmetric.

6. The EE model on 3-taxon trees.

By Definition 5.1, the EE model on a 3-leaf rooted tree ψ^+ is the set of $\kappa \times \kappa \times \kappa$ probability tensors of the form

$$P = \tilde{P} * (M_a^{-1}, M_b^{-1}, M_c^{-1})$$

for some $\tilde{P} \in \mathrm{UE}(\psi^+)$ and invertible Markov matrices M_a, M_b, and M_c.

Because of the matrix actions, this model has a nonlinear structure. This makes it more difficult to fully characterize the model EE in terms of constraints than it was for the affine linear UE model. It also means that the optimization problem for maximum likelihood may not be a convex one, making direct use of constraints for inference more appealing than attacking the optimization problem inherent to maximum likelihood.

While determining all equality constraints satisfied by the model (i.e., generators of the ideal of model invariants) is difficult computationally, here we focus on determining some of them. We will use these in section 7 in our proof of tree identifiability under the EE model. Noting that only a few constraints are utilized in the SVDquartets method, future work should investigate whether the constraints found here are useful for rooted tree inference.

Proposition 6.1. *Let P be a tensor in the EE model on $\psi^+ = ((a,b),c)$, and let $\mathrm{Cof}(A)$ denote the matrix of cofactors of a square matrix A. Then for all $k \in [\kappa]$ the matrices*

$$Q^a_{\cdot\cdot k} = P_{+\cdot\cdot}\,\mathrm{Cof}(P_{\cdot+\cdot})^T P_{\cdot\cdot k}$$

and

$$Q^b_{\cdot\cdot k} = P_{\cdot\cdot k}\,\mathrm{Cof}(P_{+\cdot\cdot})P^T_{\cdot+\cdot}$$

are symmetric, that is,

$$Q^a_{\cdot\cdot k} = (Q^a_{\cdot\cdot k})^T \tag{6.1}$$

and

$$Q^b_{\cdot\cdot k} = (Q^b_{\cdot\cdot k})^T. \tag{6.2}$$

Proof. If P is in the EE model, then $P = \tilde{P} * (M_a^{-1}, M_b^{-1}, M_c^{-1})$, with $\tilde{P} \in$ UE and M_a, M_b, M_c Markov. Then

$$P_{\cdot+\cdot} = M_a^{-T}\tilde{P}_{\cdot+\cdot}M_c^{-1} \quad \text{and} \quad P_{+\cdot\cdot} = M_b^{-T}\tilde{P}_{+\cdot\cdot}M_c^{-1} = M_b^{-T}\tilde{P}_{\cdot+\cdot}M_c^{-1}$$

since $\tilde{P} \in \text{UE}$ implies $\tilde{P}_{\cdot+\cdot} = \tilde{P}_{+\cdot\cdot}$. Then, assuming necessary inverses exist,

$$P_{\cdot+\cdot}^{-T} P_{+\cdot\cdot}^{T} = M_a M_b^{-1}.$$

Thus

$$P *_a (P_{\cdot+\cdot}^{-T} P_{+\cdot\cdot}^{T}) = \tilde{P} * (M_b^{-1}, M_b^{-1}, M_c^{-1}).$$

Note every slice of $\tilde{P} * (M_b^{-1}, M_b^{-1}, M_c^{-1})$ with fixed c-index is symmetric, since that is true for $\tilde{P}$. Thus

$$(P_{\cdot+\cdot}^{-T} P_{+\cdot\cdot}^{T})^T P_{\cdot\cdot k} = P_{+\cdot\cdot} P_{\cdot+\cdot}^{-1} P_{\cdot\cdot k}$$

is symmetric for every k. Using the cofactor formula for the inverse of a matrix and clearing denominators by multiplying by a determinant yields (6.1).

The assumption of invertibility used in this argument can be justified for a dense set of choices of $\tilde{P}$. Indeed, it is enough to exhibit one such choice, since that indicates the subset leading to noninvertibility is a proper subvariety (defined by certain minors vanishing), and hence of lower dimension. Such a choice is obtained with the Markov matrices being the identity and $\tilde{P}$ having nonzero diagonal entries and zero elsewhere. Since the claim is established on a dense set, it holds everywhere by continuity.

The claim (6.2) can be shown either in a similar way or by conjugating (in the sense of multiplying by a matrix and its transpose) $Q^a_{\cdot\cdot k}$ by $P_{\cdot+\cdot} P_{+\cdot\cdot}^{-1}$ and removing determinant factors. ∎

Remark 6.2. Since $Q^a_{\cdot\cdot k}$ and $Q^b_{\cdot\cdot k}$ are conjugate for any tensor P (even one not in the EE model), checking that one is symmetric implies the other is as well, provided the appropriate inverse exists. If these are used as necessary conditions for membership in the model, when applied to data it may still be desirable to check that both are approximately symmetric, since it is unclear how conjugation will affect the way we measure the inevitable stochastic error leading to violation of perfect symmetry.

Corollary 6.3. *The EE model on $\psi^+ = ((a,b),c)$ is contained in the algebraic variety defined by the degree $\kappa+1$ polynomials given by the entries of the 2κ matrix equations*

$$P_{+\cdot\cdot} \operatorname{Cof}(P_{\cdot+\cdot})^T P_{\cdot\cdot k} - P_{\cdot\cdot k}^T \operatorname{Cof}(P_{\cdot+\cdot}) P_{+\cdot\cdot}^T,$$
$$P_{\cdot\cdot k} \operatorname{Cof}(P_{+\cdot\cdot}) P_{\cdot+\cdot}^T - P_{\cdot+\cdot} \operatorname{Cof}(P_{+\cdot\cdot})^T P_{\cdot\cdot k}^T.$$

The polynomials of this corollary also arise as phylogenetic invariants for the general Markov (GM) model of sequence evolution [3], which independently assigns Markov matrices to each edge of the tree to describe the full substitution process along that edge. Note that the GM model is quite different from the EE model, as it has no notion of rate matrix or edge length, nor does it include a coalescent process.

In the setting of [3], the tensors of interest are those in the orbits of three-way diagonal tensors under actions of the general liner group GL_κ in each index, while here they are the orbits of tensors symmetric in two indices under the same GL_κ actions. Since diagonal tensors display this symmetry, the invariants above must also apply to the GM model. However, the

GM model on a 3-taxon tree has additional invariants of this form, for every pair of taxa, not just those in the cherry.

Remark 6.4. Using the computational algebra software Singular [7], we are able to show that for $\kappa = 2$, there are no nontrivial polynomials vanishing on the EE model. Thus, the polynomial invariants implied by Corollary 6.3 are identically zero. For $\kappa = 3$, we verified computationally that these invariants are not identically zero.

As demonstrated by methods such as SVDquartets, reframing model constraints in terms of rank conditions can be useful for developing practical methods of phylogenetic inference. With this in mind, we can reinterpret the results of Corollary 6.3 as rank conditions for the EE model. To do so, we use the following lemma, which follows a construction of G. Ottaviani that was suggested to us by L. Oeding.

Lemma 6.5. *Let A, B, C, D, E, F be six $\kappa \times \kappa$ matrices, with B, E invertible. Then*

$$CB^{-1}A + DE^{-1}F = 0$$

if and only if the $3\kappa \times 3\kappa$ matrix

$$\begin{pmatrix} 0 & A & B \\ D & 0 & C \\ E & F & 0 \end{pmatrix}$$

has rank 2κ.

Proof. Observe

$$\begin{pmatrix} 0 & A & B \\ D & 0 & C \\ E & F & 0 \end{pmatrix} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & D \\ 0 & 0 & E \end{pmatrix} \begin{pmatrix} 0 & 0 & I \\ 0 & -(CB^{-1}A + DE^{-1}F) & CB^{-1} \\ I & E^{-1}F & 0 \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & A & B \end{pmatrix}.$$

■

Corollary 6.6. *Tensors in the EE model on $\psi^+ = ((a,b),c)$ are contained in the algebraic variety defined by the degree $2\kappa + 1$ polynomials given by the $(2\kappa+1) \times (2\kappa+1)$ minors of each of the 2κ matrices*

$$\begin{pmatrix} 0 & P_{\cdot\cdot k} & P_{\cdot +\cdot} \\ -P_{\cdot\cdot k}^T & 0 & P_{+\cdot\cdot} \\ -P_{\cdot +\cdot}^T & -P_{+\cdot\cdot}^T & 0 \end{pmatrix}$$

and

$$\begin{pmatrix} 0 & P_{\cdot +\cdot}^T & P_{+\cdot\cdot}^T \\ -P_{\cdot +\cdot} & 0 & P_{\cdot\cdot k} \\ -P_{+\cdot\cdot} & -P_{\cdot\cdot k}^T & 0 \end{pmatrix}.$$

Proof. Choosing A, B, C, D, E, F in Lemma 6.5 as shown in these matrices makes the equation $CB^{-1}A + DE^{-1}F = 0$ express that $Q^a_{\cdot\cdot k}$ and $Q^b_{\cdot\cdot k}$ are symmetric, which was shown in Proposition 6.1. ■

The result of Corollary 6.6 allows one to formulate necessary conditions for EE model membership on the 3-taxon tree in terms of rank conditions on matrices, much as the SVDquartets method is based on rank conditions on matrices in the 4-taxon case.

7. Tree identifiability under the EE model. The EE model invariants of the previous section enable us to prove that the rooted tree topology is generically identifiable under the EE model. We establish these results for $\kappa \geq 4$, which includes the cases most relevant for phylogenetic analysis.

To establish identifiability, we use the following *non*identifiability result.

Lemma 7.1. *Consider a 2-taxon species tree $(a{:}x, b{:}(\ell - x))$ of total length ℓ rooted at distance $x \leq \ell$ from a, with constant population size N above the root, and any GTR rate matrix Q with stationary distribution π. Then the probability distribution matrix F of site patterns under the CK model is symmetric and independent of x.*

Proof. Using time-reversibility, the distribution can be expressed as

$$F = \int_{t=0}^{\infty} \mathrm{diag}(\pi) M_x M_{2t} M_{\ell - x} \mu_N(t) dt,$$

where $\mu_N(t)$ is the density function for coalescent times, and $M_z = \exp(Qz)$. Since the integrand, a GTR distribution, is a symmetric matrix, then so is F. Since the M_z commute, and $M_x M_{\ell - x} = M_\ell$,

$$F = \mathrm{diag}(\pi) M_\ell \int_{t=0}^{\infty} M_{2t} \mu_N(t) dt$$

has no dependence on x. ∎

Theorem 7.2. *The rooted topological tree ψ^+ is identifiable from generic probability distributions in the $EE_\kappa(\psi^+)$ model for $\kappa \geq 4$.*

Proof. We first suppose $\kappa = 4$. For the 3-taxon trees $\phi^+ = ((a, b), c)$ and $\psi^+ = ((a, c), b)$, we show that $\mathrm{EE}(\psi^+) \cap \mathrm{EE}(\phi^+)$ has measure zero within $\mathrm{EE}(\psi^+)$. To do this, it is enough to construct one point in $\mathrm{EE}(\psi^+)$ that is not in the Zariski closure of $\mathrm{EE}(\phi^+)$, since that implies the Zariski closure of the intersection of $\mathrm{EE}(\psi^+)$ and $\mathrm{EE}(\phi^+)$ is of lower dimension than $\mathrm{EE}(\psi^+)$.

Let N be an arbitrary effective population size and let $\phi^+ = ((a{:}2, c{:}0){:}1, b{:}1)$ with distances in coalescent units (number of generations divided by $2N$). Let $\mu = 1/2N$ and define Q to be the Kimura 2-parameter (K2P) rate matrix [8]

$$\begin{pmatrix} -4 & 1 & 2 & 1 \\ 1 & -4 & 1 & 2 \\ 2 & 1 & -4 & 1 \\ 1 & 2 & 1 & -4 \end{pmatrix}$$

with equilibrium distribution $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Finally, let P be the probability tensor that arises from this choice of parameters in the CK model.

Then letting $M = \exp(2Q)$, we see that $\widetilde{P} = P * (I, M, M)$ lies in $UE(\psi^+)$, which implies that $P \in EE(\psi^+)$. To see that P does not belong to the Zariski closure of $EE(\phi^+)$ by Corollary 6.3 it suffices to show that for some k

$$P_{+\cdot\cdot} \,\mathrm{Cof}(P_{\cdot + \cdot})^T P_{\cdot\cdot k} - P^T_{\cdot\cdot k} \,\mathrm{Cof}(P_{\cdot + \cdot}) P^T_{+\cdot\cdot} \neq 0. \tag{7.1}$$

Note that $P_{+\cdot\cdot}$ and $P_{\cdot+\cdot}$ are probability distribution matrices for the same model on the 2-leaf species trees $(b{:}1, c{:}1)$ and $(a{:}2, c{:}0)$. But, by Lemma 7.1,

$$P_{+\cdot\cdot} = P_{\cdot+\cdot} = P_{+\cdot\cdot}^T = P_{\cdot+\cdot}^T,$$

so

$$P_{+\cdot\cdot} P_{\cdot+\cdot}^{-T} = P_{\cdot+\cdot}^{-1} P_{+\cdot\cdot}^T = I_4.$$

To show (7.1), it is thus enough to show that some $P_{\cdot\cdot k}$ is not symmetric. This can be verified without appealing to numerical computation: For example,

$$(P_{\cdot\cdot 1})_{12} - (P_{\cdot\cdot 1})_{21} = \frac{1}{10530} e^{-20} - \frac{1}{22230} e^{-25} - \frac{1}{20007} e^{-29}.$$

If this were zero, then multiplying by e^{29} would show e is a root of a rational polynomial, contradicting its transcendence.

Thus $\mathrm{EE}(\psi^+) \cap \mathrm{EE}(\phi^+)$ has measure zero within $\mathrm{EE}(\psi^+)$.

Interchanging taxon names then shows the intersection of any two resolved 3-taxon tree models is of measure zero within them, and thus that a generic distribution in any single 3-taxon model lies only in that 3-taxon model. This establishes the theorem for 3-taxon trees when $\kappa = 4$.

For larger trees ψ^+, each displayed rooted triple determines a measure zero subset of $\mathrm{EE}(\psi^+)$ containing all points where that rooted triple may not be identifiable from marginalizations of P to those three taxa. Since there is a finite number of such sets, for a generic $P \in \mathrm{EE}(\psi^+)$, all displayed rooted triples are identifiable, and hence so is the tree ψ^+.

For $\kappa > 4$, the proof follows by embedding the 4-state rate matrix above in the upper left corner of a κ-state GTR rate matrix and setting the remaining entries to 0. ∎

Remark 7.3. Several comments are in order about the method of proof in Theorem 7.2. First, if the matrix Q is chosen to be a Jukes–Cantor rate matrix, then one finds that the same construction of P leads to a point on which the invariants for $\mathrm{EE}(\phi^+)$ vanish. That is, P is not "sufficiently generic" to identify the rooted tree. This is explored more thoroughly in the appendix.

Second, since the argument used an instance of the CK model with a K2P rate matrix, it also establishes the following, which directly applies to models used for phylogenetic inference.

Corollary 7.4. *For $\kappa = 4$, consider any submodel of EE such that each ψ^+ has an analytic parameterization and is general enough to contain the distributions from the K2P coalescent mixture model with constant population size. Then for generic parameters of that submodel the rooted topological tree ψ^+ is identifiable.*

Note that a similar result could be obtained replacing the K2P of this statement with other models. Indeed, it is immediate for a model obtained from the K2P by permutation of nucleotides, and in general requires only reproving Theorem 7.2 using another model in place of K2P. Note, however, that the result fails for the Jukes–Cantor model, as is shown in the appendix.

Nonetheless unrooted trees are still identifiable for the CK model with JC rate matrix. To establish this, note that a probability distribution for a 4-taxon tree on taxa a, b, c, d under the

EE model has the form $P = \tilde{P} * (M_a, M_b, M_c, M_d)$ with $\tilde{P}$ in the UE model and the Markov matrices invertible. As a result, its flattenings can be expressed as

$$\begin{aligned}\mathrm{Flat}_{ab|cd}(P) &= (M_a \otimes M_b)^T \mathrm{Flat}_{ab|cd}(\tilde{P})(M_c \otimes M_d),\\ \mathrm{Flat}_{ac|bd}(P) &= (M_a \otimes M_c)^T \mathrm{Flat}_{ac|bd}(\tilde{P})(M_b \otimes M_d),\\ \mathrm{Flat}_{ad|bc}(P) &= (M_a \otimes M_d)^T \mathrm{Flat}_{ad|bc}(\tilde{P})(M_b \otimes M_c).\end{aligned}$$

Since M_a, M_b, M_c, and M_d have full rank, this implies the rank of each flattening of P is equal to the rank of the corresponding flattening of $\tilde{P}$. It is then straightforward to obtain the following analogue of Theorem 3.5.

Theorem 7.5. *The SVDquartets method, using an exact method to construct a tree from a collection of quartets, gives a statistically consistent unrooted species tree topology estimator for generic parameters under the EE model, and under any submodel with an analytic parameterization general enough to contain the CK model with JC rate matrix.*

Appendix A. Pseudoexchangeability for the Jukes–Cantor model. The proof of Theorem 7.1, on the generic identifiability of the tree topology under the EE model, used a particular point in the EE model arising from the CK K2P model. Here, we show that it is not possible to use similar arguments with a point in the CK Jukes–Cantor model. We do this by specifically considering the CK Jukes–Cantor model and showing that the model always has "extra symmetries" that prevent the identification of the rooted triple tree by these invariants.

Proposition A.1. *Consider the CK Jukes–Cantor model on the tree*

$$((a{:}\ell_a, b{:}\ell_b){:}\ell_{ab}, c{:}\ell_c).$$

If $\ell_a = \ell_{ab} + \ell_c$ then the resulting probability tensor $P = (p_{ijk})$ exhibits a, c exchangeability, that is, $p_{ijk} = p_{kji}$.

Proof. Let $P = (p_{ijk})$ be a probability tensor from the CK Jukes–Cantor model on a 3-leaf tree. While P has 64 entries, because the site substitution model is the Jukes–Cantor model, it has at most five distinct entries. Thus, we may group the coordinates of P into five equivalence classes, which we represent by

$$[p_{AAA}], [p_{AAC}], [p_{ACA}], [p_{ACC}], [p_{ACG}].$$

For any representative of the equivalence class $[p_{AAA}]$, $[p_{ACA}]$, or $[p_{ACG}]$, swapping the first and third indices produces another representative of the same equivalence class. However, for representatives of the equivalence class $[p_{AAC}]$, swapping the first and third indices produces a representative of the equivalence class $[p_{ACC}]$, and vice versa. Therefore, to prove the proposition, it suffices to show that for P, $[p_{AAC}]$ and $[p_{ACC}]$ are equal. To establish this, we prove that $p_{AAC} = p_{ACC}$.

Restricting to the leaf set $\{a, b\}$, we obtain the 2-leaf rooted tree $(a{:}\ell_a, b{:}\ell_b)$ and the probability of observing state ij from the CK Jukes–Cantor model on this tree is

$$P_{ij+} = p_{ijA} + p_{ijC} + p_{ijG} + p_{ijT}.$$

Likewise, by restricting to the leaf set $\{b,c\}$, we obtain the 2-leaf rooted tree $(b{:}\ell_b+\ell_{ab}, c{:}\ell_c)$ and the probability of observing state jk from the CK Jukes–Cantor model on this tree is

$$P_{+jk} = p_{Ajk} + p_{Cjk} + p_{Gjk} + p_{Tjk}.$$

Note that since $\ell_a = \ell_{ab} + \ell_c$, the 2-leaf species trees obtained by restricting to $\{a,b\}$ and $\{b,c\}$ differ only by the location of the root. By Lemma 7.1, since the Jukes–Cantor model is a submodel of GTR, the probability distribution matrices for the JC models on these trees are symmetric and equal. Therefore, we have $P_{ij+} = P_{ji+} = P_{+ji} = P_{+ij}$. Specifically, this implies $P_{AC+} = P_{+CA}$, or

$$p_{ACA} + p_{ACC} + p_{ACG} + p_{ACT} = p_{ACA} + p_{CCA} + p_{GCA} + p_{TCA}.$$

Under the JC model, p_{ACG}, p_{ACT}, p_{GCA}, and p_{TCA} all belong to the equivalence class of coordinates with three distinct indices, which is to say, $p_{ACG} = p_{ACT} = p_{GCA} = p_{TCA}$. Thus, by cancellation, the equation above reduces to $p_{CCA} = p_{ACC}$. Since p_{CCA} and p_{AAC} are in the same JC equivalence class, this implies $p_{AAC} = p_{ACC}$. ∎

Corollary A.2. *The invariants of Corollory 6.3 associated to all of the trees $((a,b),c)$, $((a,c),b)$, and $((b,c),a)$ vanish on all probability tensors P arising from the Jukes–Cantor CK model on any of these trees.*

Proof. First consider $\tilde{P}$ arising from the CK Jukes–Cantor model on the tree $((a{:}\ell, b{:}\ell){:}\ell, c{:}0)$. By the proposition, this tensor is fully symmetric, that is, invariant under any permutation of the indices, for any positive value of ℓ. It thus lies in the UE model for all three trees. Now the probability tensor P from the CK JC model on $((a{:}\ell_a, b{:}\ell_b){:}\ell, c{:}\ell_c)$, where $\ell_a, \ell_b \geq \ell$ and $\ell_c \geq 0$, can be expressed as

$$P = \tilde{P} * (M_a, M_b, M_c),$$

where M_a, M_b, M_c are Jukes–Cantor matrices for edges of length $\ell_a - \ell$, $\ell_b - \ell$, $\ell_c - \ell$, respectively. Thus P lies in the EE model for all three trees. Therefore the invariants associated to all three trees vanish on it.

Moreover, since entries of probability tensors in the EE model are parametrized by analytic functions of the edge lengths, composing these function with the invariants gives analytic functions that vanish on a full-dimensional subset of the parameter space, which must therefore be zero on the entire parameter space. Thus the invariants vanish on the model even when the terminal edge lengths do not satisfy the assumed inequalities. ∎

REFERENCES

[1] E. S. Allman, C. Long, and J. A. Rhodes, *Species tree inference from genomic sequences using the* log-det *distance*, SIAM J. Appl. Algebra Geom., 3 (2019), pp. 107–127.

[2] E. Allman and J. Rhodes, *The identifiability of tree topology for phylogenetic models, including covarion and mixture models*, J. Comput. Biol., 13 (2006), pp. 1101–1113.

[3] E. S. Allman and J. A. Rhodes, *Phylogenetic invariants for the general Markov model of sequence mutation*, Math. Biosci., 186 (2003), pp. 113–144.

[4] J. Chifman and L. Kubatko, *Quartet inference from SNP data under the coalescent model*, Bioinformatics, 30 (2014), pp. 3317–3324.
[5] J. Chifman and L. Kubatko, *Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites*, J. Theoret. Biol., 374 (2015), pp. 35–47.
[6] A. Crowl, P. Fritsch, G. Tiley, N. Lynch, T. Ranney, H. Ashrafi, and P. Manos, *A first complete phylogenomic hypothesis for diploid blueberries (Vaccinium section cyanococcus)*, Amer. J. Bot., 109 (2022), pp. 1596–1606, https://doi.org/10.1002/ajb2.16065.
[7] W. Decker, G.-M. Greuel, G. Pfister, and H. Schönemann, *Singular* 4-3-0—*A Computer Algebra System for Polynomial Computations*, http://www.singular.uni-kl.de, 2022.
[8] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2003.
[9] J. Gaither and L. Kubatko, *Hypothesis tests for phylogenetic quartets, with applications to coalescent-based species tree inference*, J. Theoret. Biol., 408 (2016), pp. 179–186.
[10] D. Jebb, Z. Huang, M. Pippel, G. M. Hughes, K. Lavrichenko, P. Devanna, S. Winkler, L. S. Jermiin, E. C. Skirmuntt, A. Katzourakis, L. Burkitt-Gray, D. A. Ray, K. A. M. Sullivan, J. G. Roscito, B. M. Kirilenko, L. M. Dávalos, A. P. Corthals, M. L. Power, G. Jones, R. D. Ransome, D. K. N. Dechmann, A. G. Locatelli, S. J. Puechmaille, O. Fedrigo, E. D. Jarvis, M. Hiller, S. C. Vernes, E. W. Myers, and E. C. Teeling, *Six reference-quality genomes reveal evolution of bat adaptations*, Nature, 583 (2020), pp. 578–584, https://doi.org/10.1038/s41586-020-2486-3.
[11] C. Long and L. Kubatko, *The effect of gene flow on coalescent-based species tree inference*, Syst. Biol., 67 (2018), pp. 770–785.
[12] C. Long and L. Kubatko, *Identifiability and reconstructibility of a modified coalescent*, Bull. Math. Biol., 81 (2019), pp. 408–430.
[13] M. Nute, J. Chou, E. Molloy, and T. Warnow, *The performance of coalescent-based species tree estimation methods under models of missing data*, BMC Genomics, 19 (2018), 286.
[14] H. Razifard, A. Ramos, A. L. Della Valle, C. Bodary, E. Goetz, E. J. Manser, X. Li, L. Zhang, S. Visa, D. Tieman, E. van der Knaap, and A. L. Caicedo, *Genomic evidence for complex domestication history of the cultivated tomato in Latin America*, Mol. Biol. Evol., 37 (2020), pp. 1118–1132, https://doi.org/10.1093/molbev/msz297.
[15] C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, New York, 2003.
[16] D. L. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4a150, 2016.
[17] Y. Tian and L. Kubatko, *Rooting phylogenetic trees under the coalescent model using site pattern probabilities*, BMC Evol. Biol., 17 (2017), 263.
[18] M. Wascher and L. Kubatko, *Consistency of SVDquartets and maximum likelihood for coalescent-based species tree estimation*, Syst. Biol., 70 (2021), pp. 33–48, https://doi.org/10.1093/sysbio/syaa039.