

# THE SPACE OF PHYLOGENETIC MIXTURES FOR EQUIVARIANT MODELS

MARTA CASANELLAS, JESÚS FERNÁNDEZ-SÁNCHEZ, AND ANNA M. KEDZIERSKA

**ABSTRACT. Background:**

The selection of an evolutionary model to best fit given molecular data is usually a heuristic choice. In his seminal book, J. Felsenstein suggested that certain linear equations satisfied by the expected probabilities of patterns observed at the leaves of a phylogenetic tree could be used for model selection. It remained an open question, however, whether these equations were sufficient to fully characterize the evolutionary model under consideration.

**Results:** Here we prove that, for most equivariant models of evolution, the space of distributions satisfying these linear equations coincides with the space of distributions arising from mixtures of trees. In other words, we prove that the evolution of an observed multiple sequence alignment can be modeled by a mixture of phylogenetic trees under an equivariant evolutionary model if and only if the distribution of patterns at its columns satisfies the linear equations mentioned above. Moreover, we provide a set of linearly independent equations defining this space of phylogenetic mixtures for each equivariant model and for any number of taxa. Lastly, we use these results to perform a study of identifiability of phylogenetic mixtures.

**Conclusions:** The space of phylogenetic mixtures under equivariant models is a linear space that fully characterizes the evolutionary model. We provide an explicit algorithm to obtain the equations defining these spaces for a number of models and taxa. Its implementation has proved to be a powerful tool for model selection.

*Keywords:* evolutionary model, equivariant model, phylogenetic mixture, identifiability.

## BACKGROUND

The principal goal of phylogenetics is to reconstruct the ancestral relationships among organisms. Most popular phylogenetic reconstruction methods are based on mathematical models describing the molecular evolution of DNA. In

---

Research of the first and second authors partially supported by Ministerio de Educación y Ciencia MTM2009-14163-C02-02.

spite of this, there exists no unified framework for model selection and the results are highly dependent on the models and methods used in the analysis (cf. [Pos01]).

In this paper we assume the Darwinian model of evolution proceeding along phylogenetic trees and address the following question: how can the data evolving under a particular model be characterized? In other words, we look for invariants of the DNA patterns which have evolved following a tree (or a mixture of trees, as we will see below) under a particular model. The answer to this question provided in this paper leads to a complete characterization of the evolutionary model and to a novel model selection tool, which is valid for any mixture of trees.

In what follows, we briefly explain the motivation for this work. It has been shown that if the evolution along a phylogenetic tree is described by a particular model, the expected probabilities of nucleotide patterns at the leaves of the tree satisfy certain equalities (see e.g. [Fel03, p.375]). Several authors (e.g. [Fel03], [FL92], [SHSE92]) pointed out that these equalities could potentially be used to test the fitness of the model of base change. The full set of equations required for viable model selection, however, was unknown. The objective of this work is to fill in this gap and to go a step further into practical application by providing an algorithm to compute the required invariants for model selection.

In this work we consider a group of equivariant models ([DK09], [CFS11]). These models are Markov processes on trees, whose transition matrices satisfy certain symmetries: the Jukes-Cantor model, the Kimura 2 and 3 parameter models, the strand symmetric model, and the general Markov model. Our first important result, Theorem 17, states that if evolution occurs according to trees (or even mixtures of trees) under these equivariant models, then the model of evolution is completely determined by the linear space defined by the aforementioned equalities. By exhaustively studying the group of symmetries of these models, we also give a straightforward combinatorial way of determining the equations of this linear space (see Theorem 22). The implementation of the algorithm producing the equations is available as a package `SPIn` ([KDGC12], [http://genome.crg.es/cgi-bin/phylo\\_mod\\_sel/AlgModelSelection.pl](http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgModelSelection.pl)), which has proved to be a successful tool in evolutionary model selection.

Our main technique consists in proving that the linear space above coincides with the space  $\mathcal{D}^{\mathcal{M}}$  of *phylogenetic mixtures* evolving under the model  $\mathcal{M}$ , i.e. the set of points that are linear combinations of points lying in the phylogenetic varieties  $CV_T^{\mathcal{M}}$  (see Preliminaries section for specific definitions). In biological words and in the stochastic context, this is the set of vectors of expected pattern frequencies for mixtures of trees evolving under the model  $\mathcal{M}$  (not necessarily the same tree topology in the mixture, and not necessarily the same transition

matrices when the tree topologies coincide). In phylogenetics, the so-called i.i.d. hypothesis (independent and identically distributed) about the sites of an alignment is prevalent in the most simple models. When the assumption “identically distributed” is replaced it by “distributed according to the same evolutionary model”, one obtains a phylogenetic mixture.

Phylogenetic mixtures are useful in modeling heterogeneous evolutionary processes, e.g. data comprising multiple genes, selected codon positions, or rate variation across sites (e.g. [SS03]). Among a plethora of applications, they are used in orthology predictions, gene and genome annotations, species tree reconstructions, and drug target identifications.

In addition to the main result, we determine the dimension of these linear spaces and use it to give an upper bound,  $h_0(n)$ , on the number of mixtures that should be used in phylogenetic reconstruction on  $n$  taxa. This relates to the so-called *identifiability* problem in phylogenetic mixtures, which can be posed as determining the conditions that guarantee that the model parameters (discrete parameters in the form of tree topologies and the continuous parameters of the root and model distributions) can be recovered from the data. Identifiability is crucial for consistency of the maximum likelihood approaches and, though extensively studied in the phylogenetic context, few results are known (see for instance [AR06a], [APRS11], [SV07], [RS12],[CH11]).

In brief, in Theorem 30 we prove that either the tree topologies or the continuous parameters are not generically identifiable for mixtures on more than  $h_0(n)$  trees under equivariant models. Here  $h_0(n)$  is the quotient of the dimension of the linear space  $\mathcal{D}^{\mathcal{M}}$  (computed in Proposition 20) by the number of free parameters of  $\mathcal{M}$  on a trivalent tree plus one. For example, for four taxa and the Jukes-Cantor model (resp. the Kimura 3-parameter model) this result proves that mixtures on three (resp. four) or more taxa are not identifiable (i.e. either the discrete or the continuous parameters cannot be fully identified). A detailed discussion on this subject is provided in the last section.

The main tools used in this work are algebraic geometry and group theory. The reader is referred to [Har92] and [Ser77] for general references on these topics.

## MAIN TEXT

**Preliminaries.** Phylogenetic trees and Markov models of evolution have been widely used in the literature. In what follows we fix the notation needed to deal with them in our setting.

Let  $n$  be a positive integer and denote by  $[n]$  the set  $\{1, 2, \dots, n\}$ . A *phylogenetic tree*  $T$  on the set of taxa  $[n]$  is a tree (i.e. a connected graph with no loops), whose  $n$  leaves are bijectively labeled by  $[n]$ . Its vertices represent species or other biological entities and its edges represent evolutionary processes between the vertices.

We allow internal vertices of any degree and if all the internal vertices are of degree 3 we say that the tree is *trivalent*. We will denote the set of vertices of  $T$  by  $N(T)$ , the set of edges by  $E(T)$ , and the set of interior nodes by  $Int(T)$ . A *rooted tree* is a tree together with a distinguished node  $r$  called the *root*. The root induces an orientation on the edges of  $T$ , whereby the root represents the common ancestor to all the species represented in the tree. If  $e$  is an edge of a rooted tree  $T$ , we write  $pa(e)$  and  $ch(e)$  for its parent vertex (origin) and its child vertex (end), respectively. Two unrooted phylogenetic trees on the set of taxa  $[n]$  are said to have the *same tree topology* if their labeled graphs have the same topology.

We fix a positive integer  $k$  and an ordered set  $B = \{b_1, b_2, \dots, b_k\}$ . For example, for most applications we take  $B = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$  to be the set of nucleotides in a DNA sequence. We may think of  $B$  as the set of states of a discrete random variable. We call  $W$  the complex vector space  $W = \langle B \rangle_{\mathbb{C}}$  spanned by  $B$ , so that  $B$  is a natural basis of  $W$ . For algebraic convenience, we usually work over the complex field and restrict to the stochastic setting when necessary. Vectors in  $W$  are thought of as probability distributions on the set of states  $B$  if their coordinates are non-negative and sum to one. In this setting the vector  $\sum c_i b_i$  means that observation  $b_i$  occurs with probability  $c_i$ . From now on, we will identify vectors in  $W$  with their coordinates in the basis  $B$  written as a column vector, e.g. we identify  $\sum_k b_k$  with the vector  $\mathbf{1} = (1, 1, \dots, 1)^t \in W$ .

In order to model molecular evolution on a phylogenetic tree  $T$ , we consider a Markov process specified by a root distribution,  $\pi \in W$ , and a collection of transition matrices,  $\mathbf{A} = (A^e)_{e \in E(T)}$ , where each  $A^e$  is a  $k \times k$ -matrix in  $End(W)$ . The matrices  $A^e$  represent the conditional probabilities of substitution between the states in  $B$  from the parent node  $pa(e)$  to the child node  $ch(e)$  of  $e$ . We adopt the convention that the matrices  $A^e$  act on  $W$  from the right, i.e. a vector  $\omega^t$  in  $pa(e)$  maps to  $\omega^t A^e$  in  $ch(e)$ .

Distinct forms of the transition matrices give rise to different evolutionary models. Using the terminology introduced above, we proceed to the definition of evolutionary models used throughout this work.

**Definition 1.** An (*algebraic*) *evolutionary model*  $\mathcal{M}$  is specified by giving a vector subspace  $W_0 \subset W$  such that  $\mathbf{1}^t \pi \neq 0$  for some  $\pi$  in  $W_0$ , together with a multiplicatively closed vector subspace  $Mod$  (for *model*) of  $M_k(\mathbb{C})$  containing the

identity matrix. We will usually denote such a model by  $\mathcal{M} = (W_0, Mod)$ . We define the *stochastic evolutionary model*  $s\mathcal{M} = (sW_0, sMod)$  associated to  $\mathcal{M}$  by taking  $sW_0 = \{\pi \in W_0 : \mathbf{1}^t \pi = 1\}$  and  $sMod = \{A \in Mod : A\mathbf{1} = \mathbf{1}\}$ . The term “stochastic” refers to the fact that, by restricting to the points in the spaces with non-negative real entries, we obtain distributions and Markov matrices. A phylogenetic tree  $T$  together with the parameters  $\pi$  and  $\mathbf{A} = (A^e)_{e \in E(T)}$  is said to *evolve under the algebraic evolutionary model*  $\mathcal{M}$  if  $\pi \in W_0$ , and all matrices  $A^e$  lie in  $Mod$ .

**Remark 2.** Note that  $sW_0$  and  $sMod$  are not vector spaces. The condition  $\mathbf{1}^t \pi \neq 0$  in the above definition means that the sum of the coordinates of  $\pi$  is not zero. Since vectors in  $sW_0$  with non-negative coordinates represent the probability distributions for the set of observations  $B$ , this condition implies no restriction from a biological point of view. Moreover, it ensures that  $W_0 \cap \{\sum_{x \in B} \pi_x = 1\}$  has dimension equal to  $\dim(W_0) - 1$ . In particular, the simplex of stochastic vectors in  $W_0$  will form a semialgebraic set of  $\langle B \rangle_{\mathbb{R}}$  of dimension equal to  $\dim(W_0) - 1$  (as expected).

**Remark 3.** The subspace  $Mod$  of substitution matrices is usually required to be multiplicatively closed (as in the definition above) so that when two evolutionary processes are concatenated, the final process is of the same kind. The importance of this requirement is the starting point of [SFSJ12], where a different approach to the definition of “evolutionary mode” is provided.

Our definition of evolutionary models includes most of the well-known evolutionary models, namely those given in [AR06b] and the *equivariant* models (see [DK09, CFS11]).

**Example 4.** Let  $G$  be a permutation group of  $B$ , that is, a group whose elements are permutations of the set  $B$ ,  $G \leq \mathfrak{S}_k$ . Given  $g \in G$ , write  $P_g$  for the  $k \times k$ -permutation matrix corresponding to  $g$ :  $(P_g)_{i,j} = 1$  if  $g(j) = i$  and 0 otherwise. The  $G$ -equivariant evolutionary model  $\mathcal{M}_G$  is defined by taking  $Mod$  equal to

$$M(G) = \{A \in M_k(\mathbb{C}) \mid P_g A P_g^{-1} = A \text{ for all } g \in G\},$$

and  $W_0 = \{\pi \in W \mid P_g \pi = \pi \text{ for all } g \in G\}$ . These subsets are vector subspaces of  $M_k(\mathbb{C})$  and  $W$ , respectively. Moreover, if  $A_1, A_2 \in M(G)$ , then

$$P_g A_1 A_2 P_g^{-1} = (P_g A_1 P_g^{-1})(P_g A_2 P_g^{-1}) = A_1 A_2,$$

and  $A_1 A_2 \in M(G)$ . Therefore, equivariant models provide a wide family of examples of algebraic evolutionary models in the sense of Definition 1. For example, if  $B = \{A, C, G, T\}$ , it can be seen that the algebraic versions of the Jukes-Cantor model [JC69], the Kimura models with 2 or 3 parameters [Kim80, Kim81], the strand symmetric model [CS05] or the general Markov model [Cha96] are instances of equivariant models:

- if  $G = \mathfrak{S}_4$ , then  $\mathcal{M}_G$  is the *algebraic Jukes-Cantor model* JC69,
- if  $G = \langle (\text{ACGT}), (\text{AG}) \rangle$ , then  $\mathcal{M}_G$  is the *algebraic Kimura 2-parameter model* K80,
- if  $G = \langle (\text{AC})(\text{GT}), (\text{AG})(\text{CT}) \rangle$ , then  $\mathcal{M}_G$  is the *algebraic Kimura 3-parameter model* K81,
- if  $G = \langle (\text{AT})(\text{CG}) \rangle$ , then  $\mathcal{M}_G$  is known as the *strand symmetric model* SSM, and
- if  $G = \langle e \rangle$ , then  $\mathcal{M}_G$  is the *general Markov model* GMM.

Given an evolutionary model  $\mathcal{M}$  and a phylogenetic tree  $T$ , we define the *space of parameters* as

$$\text{Par}_{\mathcal{M}}(T) = W_0 \times \left( \prod_{e \in E(T)} \text{Mod} \right).$$

Similarly, we define the space of *stochastic* parameters associated to  $T$  by

$$\text{Par}_{s\mathcal{M}}(T) = sW_0 \times \left( \prod_{e \in E(T)} s\text{Mod} \right).$$

Though artificial at first glance, the use of tensors in the framework that includes the distributions on the set of patterns in  $B$  at the leaves of a phylogenetic tree is a natural choice. Indeed, if  $p_{\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_n}$  denotes the joint probability of observing  $\mathbf{x}_1$  at leaf 1,  $\mathbf{x}_2$  at leaf 2, and so on, up to  $\mathbf{x}_n$  at leaf  $n$ , then the vector  $p = (p_{b_1 \dots b_1}, p_{b_1 b_1 \dots b_2}, \dots, p_{b_k \dots b_k})$  provides a distribution on the set of patterns in  $B$  at the leaves of  $T$ , and this can be regarded as the tensor having these coordinates in the natural basis,

$$p = \sum_{\mathbf{x}_1 \dots \mathbf{x}_n \in B} p_{\mathbf{x}_1 \dots \mathbf{x}_n} \mathbf{x}_1 \otimes \dots \otimes \mathbf{x}_n.$$

This motivates the following definition.

**Definition 5.** Given a phylogenetic tree  $T$  on the set of taxa  $[n]$ , an  $[n]$ -*tensor* is any element of the tensor power

$$\mathcal{L} := \otimes_{[n]} W.$$

Given an algebraic evolutionary model  $\mathcal{M}$  and a phylogenetic tree  $T$  with root  $r$ , every Markov process on  $T$  (specified by a collection of parameters  $\pi$  and  $\mathbf{A} = (A_e)_{e \in E(T)}$ ) gives rise to a tensor in  $\mathcal{L}$  in the following way: we consider a *parametrization*

$$(1) \quad \Psi_T^{\mathcal{M}} : \text{Par}_{\mathcal{M}}(T) \longrightarrow \mathcal{L}$$

defined by

$$\Psi_T^{\mathcal{M}}(\pi, \mathbf{A}) = \sum_{\mathbf{x}_i \in B} p_{\mathbf{x}_1 \dots \mathbf{x}_n} \mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_n,$$

where

$$(2) \quad p_{\mathbf{x}_1 \dots \mathbf{x}_n} = \sum_{\mathbf{x}_v \in B, v \in \text{Int}(T)} \pi_{\mathbf{x}_r} \prod_{e \in E(T)} A_{\mathbf{x}_{pa(e)}, \mathbf{x}_{ch(e)}}^e,$$

$\mathbf{x}_v$  denotes the state at the vertex  $v$ ,  $pa(e)$  (resp.  $ch(e)$ ) is the parent (resp. child) node of  $e$ , and  $\pi_{\mathbf{x}}, \mathbf{x} \in B$ , are the coordinates of  $\pi$ . When restricted to the stochastic matrices and distributions in  $W_0$ , this parametrization corresponds to the hidden Markov process on the tree  $T$  (the leaves correspond to the observed random variables and the interior nodes to the hidden variables).

The parametrization (1) restricts to another polynomial map  $\phi_T^{\mathcal{M}} : \text{Par}_{s\mathcal{M}}(T) \rightarrow H$ , where  $H \subset \mathcal{L}$  is the hyperplane defined by  $H = \left\{ p \in \mathcal{L} \mid \sum_{\mathbf{x}_1, \dots, \mathbf{x}_n \in B} p_{\mathbf{x}_1 \dots \mathbf{x}_n} = 1 \right\}$ . Because we work in the algebraic setting, the use of the word “stochastic” in this paper is more general than usual, as we only request entries summing to one.

From now on, we will refer to this restriction as the *stochastic parametrization*  $\phi_T^{\mathcal{M}}$ . It is important to note that when we consider the distributions in  $sW_0$  and the Markov matrices in  $sMod$ , its image by  $\phi_T^{\mathcal{M}}$  lies in the standard simplex in  $\mathcal{L}$  (and thus in  $H$ ). This in turn implies that the whole image  $\text{Im } \phi_T^{\mathcal{M}}$  is contained in  $H$ .

We proceed to define the algebraic varieties associated to the parametrization maps defined above. Roughly speaking, algebraic varieties are sets of solutions to systems of polynomial equations (e.g. [Har92]).

**Definition 6.** The *stochastic phylogenetic variety*  $V_T^{\mathcal{M}}$  associated to a phylogenetic tree  $T$  is the smallest algebraic variety containing  $\text{Im } \phi_T^{\mathcal{M}} = \left\{ \phi_T^{\mathcal{M}}(\pi_r, \mathbf{A}) : (\pi_r, \mathbf{A}) \in \text{Par}_{s\mathcal{M}}(T) \right\}$  (in particular,  $V_T^{\mathcal{M}} \subset H$ ).

Similarly, the *phylogenetic variety*  $CV_T^{\mathcal{M}}$  associated to  $T$  is the smallest algebraic variety in  $\mathcal{L}$  that contains  $\text{Im } \Psi_T^{\mathcal{M}} = \left\{ \Psi_T^{\mathcal{M}}(\pi_r, \mathbf{A}) : (\pi_r, \mathbf{A}) \in \text{Par}_{\mathcal{M}}(T) \right\}$ .

Below we explain the reason for the notation of  $CV_T^{\mathcal{M}}$ , which was adopted from [AR08b].

The reader may note that the position of the root  $r$  of  $T$  played a role in the above parameterizations. It can be shown, however, that under certain mild assumptions,  $\text{Im } \Psi_T^{\mathcal{M}}$  and  $\text{Im } \phi_T^{\mathcal{M}}$  are independent of the root position in the following sense: if two phylogenetic trees have the same topology as unrooted trees, then the smallest algebraic varieties containing the corresponding image sets

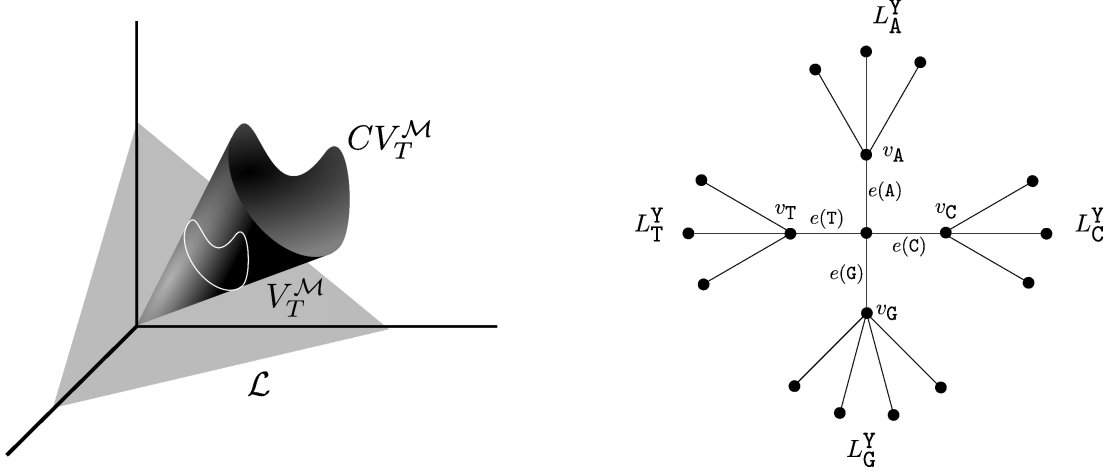


FIGURE 1. On the left, the varieties  $V_T^{\mathcal{M}}$  and  $CV_T^{\mathcal{M}}$  are shown; on the right, the phylogenetic tree described in the proof of Proposition 13 is represented.

are the same. For example, any model  $\mathcal{M} = (W_0, Mod)$  satisfying (i)  $\tilde{\pi}^t := \pi^t A$  belongs to  $W_0$  for all  $\pi \in W_0$  and all  $A \in Mod$ , and (ii)  $D_{\tilde{\pi}}^{-1} A^t D_{\pi} \in Mod$  whenever  $D_{\tilde{\pi}}^{-1}$  exists (here  $D_{\omega}$  denotes the diagonal matrix with the entries of  $\omega$  on the diagonal and zeros elsewhere) has this property (in this case, we say the model is *root-independent*). It is not difficult to check that the equivariant models satisfy these two properties (e.g. adapting the proof of [AR03] or [SHP98]). For technical reasons, from now on we consider only the evolutionary models satisfying (i) and (ii). Indeed, in this case the notation  $CV_T^{\mathcal{M}}$  refers to the fact that the phylogenetic variety is just the cone over the stochastic phylogenetic variety (see Figure 1 and the remark below).

**Remark 7.** Let  $\mathcal{M}$  be an evolutionary model satisfying (i) and (ii) above. For  $p \in \mathcal{L}$ ,  $p = \sum p_{\mathbf{x}_1 \dots \mathbf{x}_n} \mathbf{x}_1 \otimes \dots \otimes \mathbf{x}_n$ , define  $\lambda(p) := \sum_{\mathbf{x}_i \in B} p_{\mathbf{x}_1 \dots \mathbf{x}_n}$ . Then

$$CV_T^{\mathcal{M}} = \{p \in \mathcal{L} \mid p = \lambda(p)q, q \in V_T^{\mathcal{M}}\}$$

and  $V_T^{\mathcal{M}} = CV_T^{\mathcal{M}} \cap H$ . This is well known for the general Markov model [AR08b] and can be easily generalized to any model satisfying (i) and (ii).

*The space of phylogenetic mixtures.* In phylogenetics, the hypothesis that the sites of an alignment are independent and identically distributed is often used. When the assumption “identically distributed” is replaced by “distributed according to the same evolutionary model”, one obtains a phylogenetic mixture. Below, we introduce phylogenetic mixtures from the algebraic point of view (see also [MMS08]).



**Definition 8.** Fix a set of taxa  $[n]$  and an algebraic evolutionary model  $\mathcal{M}$ . A *phylogenetic mixture (on  $m$ -classes)* or  *$m$ -mixture* is any vector  $p \in \mathcal{L} = \otimes_{[n]} W$  of the form

$$p = \sum_{i=1}^m \alpha_i p^i,$$

where  $\alpha_i \in \mathbb{C}$  and  $p^i \in \text{Im}(\Psi_{T_i}^{\mathcal{M}})$  for some tree topologies  $T_i$  on the set of taxa  $[n]$ . As  $\Psi_{T_i}^{\mathcal{M}}$  is a homogeneous map, phylogenetic mixtures are represented by vectors of the form  $\sum_{i=1}^m \check{p}^i$ , where  $\check{p}^i \in \text{Im}(\Psi_{T_i}^{\mathcal{M}})$ . We call  $\mathcal{D}_{\mathcal{M}} \subset \mathcal{L}$  the *space of all phylogenetic mixtures* (on any number of classes) under the algebraic evolutionary model  $\mathcal{M}$ .

As mentioned in the introduction, the tree topologies contained in the mixture can be the same or different. An example of a phylogenetic mixture is the data modeled by the discrete Gamma-rates models (see e.g. [SS03]).

Restricting matrix rows to sum to one requires restricting the phylogenetic mixtures to the points of the form

$$q = \sum_{i=1}^m \alpha_i q^i \quad \text{where} \quad q^i \in \text{Im}(\phi_{T_i}^{\mathcal{M}}), \text{ and } \sum_i \alpha_i = 1.$$

We call  $\mathcal{D}_{s\mathcal{M}}$  the space of *stochastic phylogenetic mixtures*.

**Remark 9.** The phylogenetic variety of a trivalent tree topology contains all phylogenetic varieties of the non-trivalent tree topologies obtained by contracting any of its interior edges. Indeed, the latter are a particular case of the former when the matrices associated to the contracted edges are equal to the identity matrix. It follows that the space of phylogenetic mixtures on the trivalent tree topologies coincides with the space of phylogenetic mixtures on all possible topologies.

The following result was proven by Matsen, Mossel and Steel in [MMS08] for the two state random cluster model but, as proved below, it can be easily generalized to any evolutionary model.

**Lemma 10.** *Given a set of taxa  $[n]$  and an algebraic evolutionary model  $\mathcal{M}$ , the set of all phylogenetic mixtures  $\mathcal{D}_{\mathcal{M}}$  is a vector subspace of  $\mathcal{L}$ . Similarly,  $\mathcal{D}_{s\mathcal{M}}$  is a linear variety and it equals  $\mathcal{D}_{\mathcal{M}} \cap H$ .*

*Proof.*  $\mathcal{D}_{\mathcal{M}}$  is a  $\mathbb{C}$ -vector space and  $\mathcal{D}_{s\mathcal{M}}$  is a linear variety by their definition. It follows that  $\mathcal{D}_{\mathcal{M}}$  is an algebraic variety that contains  $\text{Im} \Psi_T^{\mathcal{M}}$  for any phylogenetic tree  $T$  on the set of taxa  $[n]$ . Therefore, it also contains  $CV_T^{\mathcal{M}}$ , and  $\mathcal{D}_{\mathcal{M}}$  equals the set of points of the form  $p = \sum p_i$ , where  $p_i \in CV_{T_i}^{\mathcal{M}}$ . Similarly,  $\mathcal{D}_{s\mathcal{M}}$  is an algebraic variety that contains  $\text{Im} \phi_T^{\mathcal{M}}$ , so it also contains  $V_T^{\mathcal{M}}$  for any phylogenetic

tree  $T$ . It follows that  $\mathcal{D}_{s\mathcal{M}}$  is formed by points of type  $q = \sum \alpha_i q_i$ , where  $q_i \in V_{T_i}^{\mathcal{M}}$  and  $\sum_i \alpha_i = 1$ .

Now we check that  $\mathcal{D}_{s\mathcal{M}} = \mathcal{D}_{\mathcal{M}} \cap H$ . Let  $q \in \mathcal{D}_{s\mathcal{M}}$ , so that  $q = \sum_{i=1}^m \alpha_i q^i$  for some  $m$ ,  $q^i \in V_{T_i}^{\mathcal{M}}$ , and  $\sum \alpha_i = 1$ . Clearly,  $q \in \mathcal{D}_{\mathcal{M}}$ . Moreover, the sum of coordinates of  $q$ ,  $\lambda(q)$ , satisfies  $\lambda(q) = \sum_i \alpha_i \lambda(q^i) = \sum_i \alpha_i = 1$ . Thus,  $q \in H$ . Conversely, let  $p = \sum_{i=1}^m p^i$  with  $p^i \in CV_{T_i}^{\mathcal{M}}$  for certain tree topologies  $T_i$ , and assume that  $\lambda(p) = 1$ . Apply Remark 7 to each  $p^i$  to get  $p^i = \lambda(p^i)q_i$  for some  $q_i \in V_{T_i}^{\mathcal{M}}$ . Then

$$p = \sum_i p^i = \sum_i \lambda(p^i)q_i$$

and  $1 = \lambda(p) = \sum_i \lambda(p^i)\lambda(q_i) = \sum_i \lambda(p^i)$  since each  $q_i$  lies on  $H$ . This proves that  $p \in \mathcal{D}_{s\mathcal{M}}$ .  $\square$

**Remark 11.** In the proof of the above lemma, we have seen that  $\mathcal{D}_{\mathcal{M}}$  and  $\mathcal{D}_{s\mathcal{M}}$  can be alternatively described as the spaces of mixtures obtained from the respective varieties  $CV_T^{\mathcal{M}}$  and  $V_T^{\mathcal{M}}$  (i.e., not only from the images of the parametrization maps).

**The space of phylogenetic mixtures for equivariant evolutionary models.** This section provides a precise description of the space  $\mathcal{D}_{\mathcal{M}}$  for the equivariant models  $\mathcal{M}$  listed in Example 4 (JC69, K80, K81, SSM, and GMM). First, we recall some definitions and facts of group theory and linear representation theory. From now on,  $B = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ ,  $k = 4$ ,  $W = \langle B \rangle_{\mathbb{C}}$ ,  $n$  is fixed and  $\mathcal{L} = \otimes_{[n]} W$ .

*Background on representation theory.* We introduce some tools in group representation theory needed in the sequel. We refer the reader to [Ser77] as a classical reference for these concepts. Although some of the following results are valid for any permutation group, for simplicity in the exposition we restrict to permutations of four elements (as our applications deal only with the case  $B = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ ).

Let  $G \leq \mathfrak{S}_4$  be a permutation group. The trivial element in  $\mathfrak{S}_4$  will be denoted as  $e$ . We write  $\rho_G$  for the restriction to  $G$  of the *defining* representation  $\rho : \mathfrak{S}_4 \rightarrow GL(W)$  given by the permutations of the basis  $B$  of  $W$ . This representation induces a  $G$ -module structure on  $W$  by setting  $g \cdot \mathbf{x} := \rho(g)(\mathbf{x}) \in W$ . In fact,  $\rho$  induces a  $G$ -module structure on any tensor power  $\otimes^s W$  by setting

$$(3) \quad g \cdot (\mathbf{x}_1 \otimes \dots \otimes \mathbf{x}_s) := g \cdot \mathbf{x}_1 \otimes \dots \otimes g \cdot \mathbf{x}_s,$$

and extending by linearity. From now on, the space  $\mathcal{L} = \otimes^n W$  will be implicitly considered as a  $G$ -module with this action. We call  $\chi$  the character associated to the representation  $\rho_G : G \rightarrow GL(W)$ , i.e.  $\chi(g)$  is the trace of the corresponding permutation matrix or, in other words,  $\chi(g)$  equals the number of fixed elements

in  $B$  by the permutation  $g \in G$ . Then the character associated to the induced representation  $G \rightarrow GL(\otimes^n W)$  is  $\chi^n$ , the  $n$ -th power of  $\chi$ .

We write  $N_1, \dots, N_t$  for the irreducible representations of  $G$  and  $\omega_1, \dots, \omega_t$  for the corresponding irreducible characters, where  $N_1$  and  $\omega_1$  will denote the trivial representation and trivial character, respectively. Maschke's Theorem applied to the action of  $G$  described in (3) states that there is a decomposition of  $\otimes^s W$  into its isotypic components:

$$(4) \quad \otimes^s W = \bigoplus_{i=1}^t (\otimes^s W)[\omega_i],$$

where each  $(\otimes^s W)[\omega_i]$  is isomorphic to a number of copies of the irreducible representation  $N_i$  associated to  $\omega_i$ ,  $(\otimes^s W)[\omega_i] \cong N_i \otimes \mathbb{C}^{m_i(s)}$ , for some non-negative integer  $m_i(s)$  called the *multiplicity of  $\otimes^s W$  relative to  $\omega_i$* . The isotypic component of  $\mathcal{L}$  associated to the trivial representation will be denoted by  $\mathcal{L}^G$  and it is composed of the  $[n]$ -tensors invariant under the action of  $G$  defined in (3). If  $\mathcal{M}$  is the equivariant evolutionary model associated to  $G$ ,  $\mathcal{L}^G$  will also be denoted as  $\mathcal{L}^{\mathcal{M}}$ . It is easy to prove that  $CV_T^{\mathcal{M}} \subset \mathcal{L}^G$  (see Lemma 4.3 of [DK09]).

We recall that the set  $\Omega_G = \{\omega_i\}_{i=1, \dots, t}$  of irreducible characters of  $G$  forms an orthonormal basis of the space of characters relative to the inner product defined by

$$(5) \quad \langle f, h \rangle := \frac{1}{|G|} \sum_{g \in G} f(g) \overline{h(g)}.$$

We introduce the following notion.

**Definition 12.** An  $n$ -word over  $B$  is an ordered sequence  $\mathbf{X} = \mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_n$ , where every letter is taken from the alphabet  $B$ . The set of  $n$ -words is equivalent to the cartesian power  $B^n$  and will be denoted by  $\mathcal{B}$ .

Words will be denoted in typewriter uppercase font (like  $\mathbf{X}$ ) and their letters in lowercase (like  $\mathbf{x}$ ). Sometimes it will be convenient to identify the  $[n]$ -tensors of the form  $\mathbf{x}_1 \otimes \dots \otimes \mathbf{x}_n$  with the  $n$ -words  $\mathbf{X} = \mathbf{x}_1 \dots \mathbf{x}_n$ . Consequently, we will identify  $\mathcal{B}$  with the natural basis of  $\mathcal{L}$ . Given  $\mathbf{X} \in \mathcal{B}$ , we will denote by  $\{\mathbf{X}\}_G = \{g\mathbf{X} \mid g \in G\}$  the  $G$ -orbit of  $\mathbf{X}$ . We associate a  $G$ -invariant tensor,  $\tau\{\mathbf{X}\}_G$ , to each orbit  $\{\mathbf{X}\}_G$ :  $\tau\{\mathbf{X}\}_G := \sum_{g \in G} g\mathbf{X}$ . It is straightforward to see that every  $G$ -invariant tensor can be written as a linear combination of the tensors  $\tau\{\mathbf{X}\}_G$ ,  $\mathbf{X} \in \mathcal{B}$ . On the other hand, the set of different  $\tau\{\mathbf{X}\}_G$ 's is linearly independent, since the corresponding  $G$ -orbits  $\{\mathbf{X}\}_G$  have non-overlapping composition of the elements of  $\mathcal{B}$ .

*Mixtures for equivariant models.* For each  $\mathbf{x} \in B$ , we write  $S_G(\mathbf{x})$  for the stabiliser of  $\mathbf{x}$  under the action of  $G$ , that is,  $S_G(\mathbf{x}) = \{g \in G : g \cdot \mathbf{x} = \mathbf{x}\}$ .

**Proposition 13.** *Let  $G$  be a subgroup of  $\mathfrak{S}_4$  such that  $S_G(\mathbf{x}_0) = \{e\}$  for some  $\mathbf{x}_0 \in B$ . Then every tensor of type  $\tau\{\mathbf{X}\}_G$ ,  $\mathbf{X} \in \mathcal{B}$ , lies in the image of  $\Psi_T^{\mathcal{M}_G}$  for some tree topology  $T$ . In particular,  $\mathcal{L}^G \subset \mathcal{D}_{\mathcal{M}_G}$ .*

*Proof.* For any  $G$ -orbit  $\{\mathbf{Y}\}_G$ ,  $\mathbf{Y} \in \mathcal{B}$ , write  $\tau\{\mathbf{Y}\}_G = y_1 \otimes \dots \otimes y_n + \sum_{g \neq e} g \cdot y_1 \otimes \dots \otimes g \cdot y_n$ . We will explicitly associate a tree topology and parameters  $(\pi, \mathbf{A})$  to it so that the tensor  $\tau\{\mathbf{T}\}_G$  is equal to  $\Psi_T^{\mathcal{M}_G}$ . To this aim, we denote by  $B(\mathbf{Y})$  the set of letters appearing in  $\mathbf{Y}$ . Then for every  $\mathbf{z} \in B(\mathbf{Y})$ , consider the set  $L_{\mathbf{z}}^{\mathbf{Y}} = \{i \in [n] : y_i = \mathbf{z}\}$ , so that  $\cup_{\mathbf{z} \in B(\mathbf{Y})} L_{\mathbf{z}}^{\mathbf{Y}} = [n]$ .

We construct a tree  $T$  on the set of taxa  $[n]$  in the following way. We join each taxa in  $L_{\mathbf{z}}^{\mathbf{Y}}$  to a common node  $v_{\mathbf{z}}$  by an edge. Then each vertex  $v_{\mathbf{z}}$  is joined to the root of the tree (we call it  $r$ ) by an edge that we denote as  $e(\mathbf{z})$  (see figure 1). Now, in the edges joining any  $v_{\mathbf{z}}$  with some leaf in  $L_{\mathbf{z}}^{\mathbf{Y}}$ , we consider the identity matrix, while the matrix in  $e(\mathbf{z})$  is defined by taking

$$A_{i,j}^{e(\mathbf{z})} = \begin{cases} 1 & \text{if } (i, j) = (h \cdot \mathbf{x}_0, h \cdot \mathbf{z}) \text{ for some } h \in G, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, if  $c$  is the cardinality of  $\{\mathbf{x}_0\}_G$ , define the distribution at the root  $\pi = (\pi_{\mathbf{A}}, \pi_{\mathbf{C}}, \pi_{\mathbf{G}}, \pi_{\mathbf{T}})$  by

$$\pi_{\mathbf{z}} = \begin{cases} \frac{1}{c} & \text{if } \mathbf{z} \in \{\mathbf{x}_0\}_G, \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to check that these matrices and the vector  $\pi$  are  $G$ -equivariant, so  $(\pi, \mathbf{A}) \in \text{Par}_{\mathcal{M}_G}(T)$ . Now, from (2) and the definition of  $\pi$ , we can write

$$P_{\mathbf{x}_1 \dots \mathbf{x}_n} = \sum_{\substack{g \in G \\ \{\mathbf{x}_{\mathbf{z}}\}_{\mathbf{z} \in B(\mathbf{Y})} \subset B}} P_{\mathbf{x}_1 \dots \mathbf{x}_n}(g, \{\mathbf{x}_{\mathbf{z}}\}_{\mathbf{z} \in B(\mathbf{Y})})$$

where

$$P_{\mathbf{x}_1 \dots \mathbf{x}_n}(g, \{\mathbf{x}_{\mathbf{z}}\}_{\mathbf{z} \in B(\mathbf{Y})}) = \pi_{g \cdot \mathbf{x}_0} \prod_{\mathbf{z} \in B(\mathbf{Y})} \left( A_{g \cdot \mathbf{x}_0, \mathbf{x}_{\mathbf{z}}}^{e(\mathbf{z})} \prod_{j \in L_{\mathbf{x}_{\mathbf{z}}}^{\mathbf{Y}}} \delta_{\mathbf{x}_{\mathbf{z}}, \mathbf{x}_j} \right)$$

(here  $\delta_{a,b}$  stands for the Kronecker delta, i.e.  $\delta_{a,a} = 1$ ,  $\delta_{a,b} = 0$  if  $a \neq b$ ). Moreover, from the definition of the matrix  $A^{e(\mathbf{z})}$ , we have

$$A_{g \cdot \mathbf{x}_0, \mathbf{x}_{\mathbf{z}}}^{e(\mathbf{z})} = \begin{cases} 1 & \text{if } (g \cdot \mathbf{x}_0, \mathbf{x}_{\mathbf{z}}) = (h \cdot \mathbf{x}_0, h \cdot \mathbf{z}) \text{ for some } h \in G, \\ 0 & \text{otherwise.} \end{cases}$$

The hypothesis  $S_G(\mathbf{x}_0) = \{e\}$  ensures that  $(g \cdot \mathbf{x}_0, \mathbf{x}_{\mathbf{z}}) = (h \cdot \mathbf{x}_0, h \cdot \mathbf{z})$  if and only if  $g = h$ . From this, it becomes clear that  $P_{\mathbf{x}_1 \dots \mathbf{x}_n}(g, \{\mathbf{x}_{\mathbf{z}}\}_{\mathbf{z} \in B(\mathbf{Y})}) = 0$  unless

- (1)  $\mathbf{x}_z = g \cdot \mathbf{z}$ , for  $\mathbf{z} \in B$ , and
- (2) for each  $i \in L_z^Y$ ,  $\mathbf{x}_i$  is equal to  $\mathbf{x}_z = g \cdot \mathbf{z}$ ,

in which case  $P_{\mathbf{x}_1 \dots \mathbf{x}_n}(g, \{\mathbf{x}_z\}_{z \in B(Y)}) = \pi_{g \cdot x_0} = \frac{1}{c}$ . It follows that

$$p_{\mathbf{x}_1 \dots \mathbf{x}_n} = \begin{cases} 1 & \text{if } \mathbf{x}_1 \dots \mathbf{x}_n \in \{Y\}_G, \\ 0 & \text{otherwise,} \end{cases}$$

and  $\Psi_T^M(\pi, \mathbf{A}) = \tau\{Y\}_G$ . Moreover, as the set of  $\tau\{Y\}_G$ , for  $Y \in \mathcal{B}$ , generates the vector space  $\mathcal{L}^G$ , the second claim follows.  $\square$

**Remark 14.** The above result is not true if the hypothesis  $S_G(\mathbf{x}_0) = \langle e \rangle$  is removed. For example, if  $G = \langle (\text{ACGT}), (\text{AG}) \rangle$  (so that  $\mathcal{M} = \text{K80}$ ), then  $S_G(\mathbf{A}) = S_G(\mathbf{G}) = \{e, (\text{CT})\}$  and  $S_G(\mathbf{C}) = S_G(\mathbf{T}) = \{e, (\text{AG})\}$ . In that case, it can be shown that the  $G$ -orbit  $\{\text{ACGT}\}_G$  is not in  $\text{Im}\Psi_T^{\text{K80}}$  for any tree topology  $T$  with 4 leaves.

Since the above condition on the group holds for  $G = \mathfrak{S}_4$ ,  $G = \langle (\text{AT})(\text{CG}) \rangle$ , and  $G = \langle (\text{AG}), (\text{ACGT}) \rangle$ , we deduce the following claim.

**Corollary 15.** *If  $G$  corresponds to any of the equivariant models K81, SSM or GMM, we have  $\mathcal{L}^G \subset \mathcal{D}_{\mathcal{M}_G}$ .*

In phylogenetics, an *invariant of a phylogenetic tree  $T$*  is an equation satisfied by the expected distributions of patterns at the leaves of  $T$ , irrespectively of the continuous parameters of the model  $\mathcal{M}$ . In the algebraic geometry setting, these are the equations satisfied by all  $p \in CV_T^M$ . Invariants were introduced by Lake (see [Lak87]) and Cavender and Felsenstein (see [CF87]). A *phylogenetic invariant of  $T$*  is an invariant of  $T$ , which is not an invariant of all other phylogenetic trees (under the same model  $\mathcal{M}$ ). Equivalently,  $f$  is a phylogenetic invariant of  $CV_T^M$  if it is an invariant of  $CV_T^M$  and there exists a tree topology  $T'$  such that  $f$  is not an invariant of  $CV_{T'}^M$ . In principle, phylogenetic invariants can be used for tree topology reconstruction purposes.

**Remark 16.** (a) It can be seen that the condition of trivial stabiliser for some element of  $B$  given in Proposition 13 guarantees that all the irreducible representations of  $G$  will be present in the decomposition of  $W$  into its isotypic components. Then, by using the results of [CFS11], it follows that the corresponding equivariant model will have no linear phylogenetic invariants. This fact was already known for the models in the above corollary: see [AR04] for the GMM, [CS05] for the SSM and [SS05] for the K81. Here we provided an alternative proof based on elementary tools of group theory.

- (b) The models JC69 and K80 are known to have linear phylogenetic invariants, but these are the only linear invariants which do not define hyperplanes containing  $\mathcal{L}^G$ , as can be deduced from [FL92, SS05]. In fact, for

these two models, the claim of the corollary is still true as stated in the following theorem. Nevertheless, we have not been able to provide a unified proof of this fact because of the different properties of the corresponding groups. There is no description of the space of linear invariants for other equivariant models not listed in Example 4, so we cannot claim that the result below still holds.

**Theorem 17.** *If  $\mathcal{M}_G$  is one of the equivariant evolutionary models JC69, K80, K81, SSM, or GMM, then the space of phylogenetic mixtures  $\mathcal{D}_{\mathcal{M}_G}$  coincides with  $\mathcal{L}^G$ , and  $\mathcal{D}_{s\mathcal{M}_G}$  equals  $\mathcal{L}^G \cap H$ .*

This theorem allows to identify the set of all phylogenetic mixtures  $\mathcal{D}_{\mathcal{M}_G}$  with  $\mathcal{L}^G$ , which is a vector subspace of  $\mathcal{L}$  whose linear equations are easy to describe. In other words,  $\mathcal{L}^G$  is the smallest linear space containing the data coming from any mixture of trees evolving under the model  $\mathcal{M}_G$ . One can therefore use  $\mathcal{L}^G$  to select the most suitable model for the given data. This has been studied in [KDGC12].

*Proof of Theorem 17.* For equivariant models we have that  $CV_T^{\mathcal{M}_G} \subset \mathcal{L}^G$  for any tree  $T$ . Hence, by Lemma 10 and the definition of  $\mathcal{D}_{\mathcal{M}_G}$ ,  $\mathcal{D}_{\mathcal{M}_G}$  is a vector subspace of  $\mathcal{L}^G$ .

From Corollary 15, we infer the equality  $\mathcal{L}^G = \mathcal{D}_{\mathcal{M}_G}$  for the models K81, SSM and GMM. For the other two models, JC69 and K80, it remains to prove that there does not exist any hyperplane  $\Pi$  containing  $\mathcal{D}_{\mathcal{M}_G}$  and not containing  $\mathcal{L}^G$ . If such a hyperplane existed, then it would contain all the points of  $CV_T^{\mathcal{M}_G}$  for any tree topology  $T$ . It suffices to prove that for these models there are no homogeneous linear polynomials vanishing on all tree topologies, except for the linear equations vanishing on  $\mathcal{L}^G$ . This has been seen in Remark 16(b).

The equality  $\mathcal{D}_{s\mathcal{M}_G} = \mathcal{L}^G \cap H$  follows immediately from Lemma 10 and the first assertion in the statement of this theorem.  $\square$

**Remark 18.** We are indebted to one of the referees of this paper for pointing out that the preceding result, as well as the second part of Proposition 13, can also be inferred from Proposition 4.9 of [DK09]: under the assumption that the stabiliser of some state is trivial, Draisma and Kuttler show that the star tree is the smallest algebraic variety containing the tensors  $\tau\{\mathbf{X}\}_G$ , for *pure* tensors  $\mathbf{X}$  (that is, tensors of rank 1). It follows that the set of mixtures on the star tree equals the space  $\mathcal{L}^G$ .

**Remark 19.** It is not difficult to check that for  $\mathcal{M} = \text{K81, SSM or GMM}$ ,  $\mathcal{D}_{\mathcal{M}}$  coincides with the space of mixtures on the star tree (see also [MMS08], where

the same result is proven for a 2-state model). On the contrary, this is not true for JC69 and K80 models because in this case the star tree lies in a smaller linear space as a consequence of the existence of phylogenetic linear equations (see Remark 16(b)).

*Equations for the space  $\mathcal{L}^G$ .* Our goal here is to compute the dimension of  $\mathcal{L}^G$  for the groups associated to the equivariant models listed in Definition 4, and to list a set of independent linear equations defining this space.

**Proposition 20.** *Using the notations above,*

- (i)  $\dim \mathcal{L}^{\text{SSM}} = 2^{2n-1}$ ,
- (ii)  $\dim \mathcal{L}^{\text{K81}} = 4^{n-1}$ ,
- (iii)  $\dim \mathcal{L}^{\text{K80}} = 2^{2n-3} + 2^{n-2}$ , and
- (iv)  $\dim \mathcal{L}^{\text{JC69}} = \frac{2^{2n-3}+1}{3} + 2^{n-2}$ .

*Proof.* Let  $\mathcal{M}$  be any equivariant model. By definition, we know that  $\mathcal{L}^G$  is the isotypic component of  $\otimes^n W$  associated to the trivial representation  $(\otimes^n W)_{[\omega_1]}$ . Since the dimension of the trivial representation is one, it follows that the dimension of  $\mathcal{L}^{\mathcal{M}}$  is precisely the multiplicity  $m_1(n)$ , i.e. the number of times the trivial representation appears in the decomposition of  $\otimes^n W$  into isotypic components. This multiplicity  $m_1(n)$  equals (see (5))

$$\langle \chi^n, \omega_1 \rangle = \frac{1}{|G|} \sum_{g \in G} \chi^n(g) \omega_1(g).$$

The proof ends by grouping the elements of  $G$  in the conjugacy classes of  $G$  for SSM, K81, K80, or JC69. Recall that the conjugacy classes of a group  $G$  are the disjoint sets of the form  $C(g) = \{h^{-1}gh : h \in G\}$ . If  $C_1, \dots, C_s$  are the conjugacy classes for  $G$ , write  $\mathcal{C}(G) = (|C_1|, \dots, |C_s|)$  for the  $s$ -tuple of their cardinalities, so that  $\sum_{i=1}^s |C_i| = |G|$ . Recall that  $\chi^n(g_1) = \chi^n(g_2)$  whenever  $g_1$  and  $g_2$  lie in the same conjugacy class, so we can represent  $\chi^n$  by an  $s$ -tuple  $\chi_{\mathcal{C}(G)}^n = (t_1, \dots, t_s)$ , where  $t_i = \chi^n(g)$  for any  $g \in C_i$ . Thus, we have  $m_1(n) = \frac{1}{|G|} \sum_{i=1}^s \chi^n(g_i) |C_i|$ , where  $g_i$  is any element in the conjugacy class  $C_i$ . The result for  $\mathcal{M} = \text{SSM}, \text{K81}, \text{K80}$ , or JC69 follows by applying the following table.

$G \leq \mathfrak{S}_4$	$\mathcal{M}$	representatives of conj. classes	$\mathcal{C}(G)$	$\chi_{\mathcal{C}(G)}^n$
$\langle (\text{AT})(\text{CG}) \rangle$	SSM	$\{e, (\text{AT})(\text{CG})\}$	(1, 1)	$(4^n, 0)$
$\langle (\text{AC})(\text{GT}), (\text{AG})(\text{CT}) \rangle$	K81	$\{e, (\text{AT})(\text{CG}), (\text{AC})(\text{GT}), (\text{AG})(\text{CT})\}$	(1, 1, 1, 1)	$(4^n, 0, 0, 0)$
$\langle (\text{ACGT}), (\text{AG}) \rangle$	K80	$\{e, (\text{AC})(\text{GT}), (\text{AG})(\text{CT}), (\text{ACGT}), (\text{AG})\}$	(1, 2, 1, 2, 2)	$(4^n, 0, 0, 0, 2^n)$
$\mathfrak{S}_4$	JC69	$\{e, (\text{AC})(\text{GT}), (\text{ACGT}), (\text{AG}), (\text{ACG})\}$	(1, 3, 6, 6, 8)	$(4^n, 0, 0, 2^n, 1)$

□

Our next goal is to provide a set of independent linear equations for  $\mathcal{L}^G$ . Before stating the main result, let us introduce some useful notation.

**Notation 21.** We consider the following subsets of  $\mathcal{B} = B^n$ :

$$\begin{aligned}\mathcal{B}_0 &= \{A \dots A, C \dots C, G \dots G, T \dots T\}, \\ \mathcal{B}_{AC|GT} &= \{A, C\}^n \cup \{G, T\}^n, \\ \mathcal{B}_{AG|CT} &= \{A, G\}^n \cup \{C, T\}^n, \\ \mathcal{B}_{AT|CG} &= \{A, T\}^n \cup \{C, G\}^n, \text{ and} \\ \mathcal{B}_2 &= \mathcal{B}_{AC|GT} \cup \mathcal{B}_{AG|CT} \cup \mathcal{B}_{AT|CG}.\end{aligned}$$

The set  $\mathcal{B}_0$  is composed of all  $n$ -words with only one letter and it is contained in  $\mathcal{B}_{AC|GT}$ ,  $\mathcal{B}_{AG|CT}$ , and  $\mathcal{B}_{AT|CG}$ . Similarly,  $\mathcal{B}_2$  is composed of all  $n$ -words with two letters at most. It is straightforward to check that  $|\mathcal{B}_{AC|GT}| = |\mathcal{B}_{AG|CT}| = |\mathcal{B}_{AT|CG}| = 2^{n+1}$  and  $|\mathcal{B}_2| = 3 \cdot 2^{n+1} - 8$ .

We adopt multiplicative notation for  $n$ -words, for instance, we write  $C^l$  for the word  $\underbrace{C \dots C}_l$ , and  $(A^l)(G^m)x_{l+m+1} \dots x_n$  for  $\underbrace{A \dots A}_l \underbrace{G \dots G}_m x_{l+m+1} \dots x_n$ , where  $x_{l+m+1}, \dots, x_n$  represent any letters.

The main result of this section is the following:

**Theorem 22.** *A set of linearly independent equations  $\mathbb{E}^{\mathcal{M}}$  defining  $\mathcal{L}^{\mathcal{M}}$  for  $\mathcal{M} = \text{JC69}, \text{K80}, \text{K81}, \text{ or SSM}$  is given by*

- $\mathbb{E}^{\text{SSM}}$  : equations  $p_X = p_{(\text{AT})(\text{CG})X}$  for all  $X \in \mathcal{B}$  with  $x_1 \in \{A, C\}$ ;
- $\mathbb{E}^{\text{K81}}$  : the equations in  $\mathbb{E}^{\text{SSM}}$ , and the equations  $\{p_X = p_{(\text{AC})(\text{GT})X}\}$  for all  $X \in \mathcal{B}$  with  $x_1 = A$ ;
- $\mathbb{E}^{\text{K80}}$  : the equations in  $\mathbb{E}^{\text{K81}}$ , plus the equations  $\{p_X = p_{(\text{AG})X}\}$  for all  $X \in \mathcal{B} \setminus \mathcal{B}_{AC|GT}$  having  $x_1 = A$  and satisfying the following condition: if  $T$  appears in  $X$ , then there is a  $C$  in a preceding position;
- $\mathbb{E}^{\text{JC69}}$  : the equations in  $\mathbb{E}^{\text{K80}}$ , together with the equations  $\{p_X = p_{(\text{AT})X}\}$  for all  $X \in \mathcal{B}_{AC|GT} \setminus \mathcal{B}_0$  of the form  $(A^l)(C^m)x_{l+m+1} \dots x_n$ ; plus the equations  $\{p_X = p_{(\text{AC})X}\}$  and  $\{p_X = p_{(\text{AT})X}\}$  for all  $X \in \mathcal{B} \setminus \mathcal{B}_2$  of the form  $(A^l)(C^m)x_{l+m+1} \dots x_n$  and satisfying the condition: if  $T$  appears in  $X$ , then there is a  $G$  in a preceding position.

The number of equations added in each case is  $2^{2n-1}$  for SSM,  $2^{2n-2}$  for K81,  $2^{2n-3} - 2^{n-2}$  for K80, and  $2^{n-1} - 1 + 2\left(\frac{2^{2n-3}+1}{3} - 2^{n-2}\right)$  for JC69.

Before proving this theorem, we explain how these sets of equations were obtained. Notice that a system of linear equations of  $\mathcal{L}^G$  is given by

$$\{p_{gX} = p_X \mid g \in G, X \in \mathcal{B}\}.$$



The role played by the  $G$ -orbits on  $\mathcal{B}$  becomes apparent. Indeed, the idea is to relate the equations to the orbits of a subgroup of  $G$ . To this aim, let  $H$  be a subgroup of  $G$  and write  $H \backslash G = \{Hg : g \in G\}$  for the set of right cosets of  $H$  in  $G$ . We consider a transversal of  $H \backslash G$ , i.e. a collection  $\{g_1, \dots, g_{[G:H]}\}$  such that  $G = \bigsqcup_{i=1}^{[G:H]} Hg_i$ . Then the orbit of any  $\mathbf{X} \in \mathcal{B}$  can be decomposed as

$$(6) \quad \{\mathbf{X}\}_G = \bigcup_{i=1, \dots, [G:H]} \{g_i \mathbf{X}\}_H.$$

This decomposition establishes the connection between the  $G$ -orbits and the  $H$ -orbits. In order to obtain a system of equations for  $\mathcal{L}^G$ , once  $\mathbb{E}^H$  has been computed, it is enough to add the equations involving the permutations in a transversal  $\{g_1 = e, g_2, \dots, g_{[G:H]}\}$  of  $H \backslash G$ :

$$\left. \begin{array}{l} p_{\mathbf{X}} = p_{g_2 \mathbf{X}} \\ p_{\mathbf{X}} = p_{g_3 \mathbf{X}} \\ \dots \\ p_{\mathbf{X}} = p_{g_{[G:H]} \mathbf{X}} \end{array} \right\} \text{for all } \mathbf{X} \in \mathcal{B}.$$

Notice that the union in (6) is not necessarily disjoint as it may happen that  $\{g_i \mathbf{X}\}_H = \{g_j \mathbf{X}\}_H$  for  $i \neq j$ . In this case, the equality  $p_{g_j \mathbf{X}} = p_{g_i \mathbf{X}}$  already holds in the space  $\mathcal{L}^H$  and does not provide any new restriction. In order to avoid this situation and obtain a minimal set of equations for  $\mathcal{L}^G$ , we request the special conditions on the  $\mathbf{X} \in \mathcal{B}$  in the statement of the theorem.

*Proof.* For each model  $\mathcal{M}$ , we prove that the corresponding equations are linearly independent and there are as many equations as the codimension of  $\mathcal{L}^{\mathcal{M}}$ . By Proposition 20, the codimension of  $\mathcal{L}^{\mathcal{M}}$  is  $2^{2n-1}$  for SSM,  $3 \cdot 4^{n-1}$  for K81,  $7 \cdot 2^{2n-3} - 2^{n-2}$  for K80, and  $4^n - \frac{2^{2n-3}+1}{3} - 2^{n-2}$  for JC69. In the sequel, we refer to the groups by the name of the equivariant model associated to them.

**SSM:** As SSM is the group  $\{e, (\text{AT})(\text{CG})\}$ , a set of equations for SSM is  $\{p_{\mathbf{X}} = p_{(\text{AT})(\text{CG})\mathbf{X}}\}$ . Fixing  $\mathbf{x}_1$  in  $\{\mathbf{A}, \mathbf{C}\}$  we obtain  $2^{2n-1}$  linearly independent equations (equations involving different coordinates). The codimension of  $\mathcal{L}^{\text{SSM}}$  is equal to  $2^{2n-1}$ , which coincides with the number of equations given, and thus this set of equations defines  $\mathcal{L}^{\text{SSM}}$ .

**K81:** Since a transversal of  $\text{SSM} \backslash \text{K81}$  is  $\{e, (\text{AC})(\text{GT})\}$ , the hyperplanes  $p_{\mathbf{X}} = p_{(\text{AC})(\text{GT})\mathbf{X}}$  contain  $\mathcal{L}^{\text{K81}}$  but not  $\mathcal{L}^{\text{SSM}}$ . Moreover, using (6) we see that the orbit  $\{\mathbf{X}\}_{\text{K81}}$  decomposes into the disjoint union of  $\{\mathbf{X}\}_{\text{SSM}}$  and  $\{(\text{AC})(\text{GT})\mathbf{X}\}_{\text{SSM}}$  for any  $\mathbf{X} \in \mathcal{B}$ . Therefore, the equations given for K81 involve different coordinates than those in  $\mathbb{E}^{\text{SSM}}$ . Requiring  $\mathbf{x}_1 = \mathbf{A}$ , we obtain  $4^{n-1}$  linearly independent new equations. Thus  $\mathbb{E}^{\text{K81}}$  defines the space  $\mathcal{L}^{\text{K81}}$  because the

number of linearly independent equations provided,  $2^{2n-1} + 4^{n-1} = 3 \cdot 4^{n-1}$ , coincides with the codimension of  $\mathcal{L}^{K81}$ .

**K80:** The set  $\{e, (AG)\}$  is a transversal of  $K81 \setminus K80$ . In order to show that the equations provided are linearly independent to those of  $\mathbb{E}^{K81}$ , we apply (6) to this transversal to obtain  $\{X\}_{K80} = \{X\}_{K81} \cup \{(AG)X\}_{K81}$ . If  $X \notin \mathcal{B}_{AG|CT}$ , then  $\{(AG)X\}_{K81}$  and  $\{X\}_{K81}$  are disjoint, so each equation  $p_X = p_{(AG)X}$  is linearly independent from  $\mathbb{E}^{K81}$ . The set  $\mathcal{B} \setminus \mathcal{B}_{AG|CT}$  has cardinal  $4^n - 2^{n+1}$  and, if  $X \in \mathcal{B} \setminus \mathcal{B}_{AG|CT}$ , each orbit  $\{X\}_{K80}$  has cardinality 8. Therefore, the number of different orbits for  $X \in \mathcal{B} \setminus \mathcal{B}_{AG|CT}$  is  $(4^n - 2^{n+1})/8 = 2^{2n-3} - 2^{n-2}$ . Moreover, the choice of  $X$ 's in  $\mathcal{B} \setminus \mathcal{B}_{AG|CT}$  with  $x_1 = A$  and satisfying "if  $T$  appears in  $X$ , there is a  $C$  in a preceding position" guarantees that we take only one element in each  $\{X\}_{K80}$ , and thus we are adding exactly one equation for each of these  $X$ 's. Overall, there are  $3 \cdot 4^{n-1} + (2^{2n-3} - 2^{n-2}) = 7 \cdot 2^{2n-3} - 2^{n-2}$  linearly independent equations in  $\mathbb{E}^{K80}$ . This number coincides with the codimension of  $\mathcal{L}^{K80}$  and these equations define  $\mathcal{L}^{K80}$ .

**JC69:** A transversal of  $K80 \setminus JC69$  is  $\{e, (AC), (AT)\}$ , therefore (6) applies to give  $\{X\}_{JC69} = \{X\}_{K80} \cup \{(AC)X\}_{K80} \cup \{(AT)X\}_{K80}$ .

- if  $X \in \mathcal{B}_{AC|GT} \setminus \mathcal{B}_0$ , then  $\{(AC)X\}_{K80} = \{X\}_{K80}$  and  $\{X\}_{JC69}$  is the disjoint union of  $\{X\}_{K80}$  and  $\{(AT)X\}_{K80}$ . As such, each equation  $p_X = p_{(AT)X}$  is linearly independent from  $\mathbb{E}^{K80}$ .

Moreover, if  $X \in \mathcal{B}_{AC|GT} \setminus \mathcal{B}_0$  is of the form  $(A^l)(C^m)x_{l+m+1} \dots x_n$ , we have  $2^{n-1} - 1$  such equations and they are linearly independent.

- if  $X \in \mathcal{B} \setminus \mathcal{B}_2$  then the three orbits  $\{(AC)X\}_{K80}$ ,  $\{(AT)X\}_{K80}$ , and  $\{X\}_{K80}$  have 8 elements each and are disjoint. Therefore, for these  $X$ 's, each equation of type  $\{p_X = p_{(AC)X}\}$  or  $\{p_X = p_{(AT)X}\}$  is linearly independent from  $\mathbb{E}^{K80}$ . Moreover, as  $\mathcal{B} \setminus \mathcal{B}_2$  has cardinal  $4^n - 3 \cdot 2^{n+1} + 8$  and is covered by these orbits, we have  $\frac{4^n - 3 \cdot 2^{n+1} + 8}{24} = \frac{1}{3}(2^{2n-3} + 1) - 2^{n-2}$  different orbits. The restriction to the elements of the form  $(A^l)(C^m)x_{l+m+1} \dots x_n$  and satisfying that "if  $T$  appears in  $X$ , there is some  $G$  in a preceding position" guarantees that the equations are written only once for each orbit.

Summing up, there are

$$7 \cdot 2^{2n-3} - 2^{n-2} + \left( 2^{n-1} - 1 + 2 \left( \frac{1}{3}(2^{2n-3} + 1) - 2^{n-2} \right) \right)$$

linearly independent equations in  $\mathbb{E}^{JC69}$  that contain  $\mathcal{L}^{JC69}$ . As this number is equal to the codimension  $4^n - \frac{2^{2n-3} + 1}{3} - 2^{n-2}$  of  $\mathcal{L}^{JC69}$ , the proof is complete.

All the equalities among orbits used in this proof are summarized in the following table (where  $\dots$  means 'the set on the left' and  $\text{"}$  means 'the set on the top').

	$\{X\}_{\text{GMM}}$	$\{X\}_{\text{SSM}}$	$\{X\}_{\text{K81}}$	$\{X\}_{\text{K80}}$	$\{X\}_{\text{JC69}}$
$\mathcal{B}_0$	$\{X\}$	$\dots \cup \{(AT)(CG)X\}$	$\dots \cup \{(AC)(GT)X\}_{\text{SSM}}$	$\dots$	$\dots$
$\mathcal{B}_{\text{AG CT}}$	"	"	"	$\dots$	$\dots \cup \{(AC)X\}_{\text{K80}}$
$\mathcal{B}_{\text{AC GT}}$	"	"	"	$\dots \cup \{(AG)X\}_{\text{K81}}$	$\dots \cup \{(AT)X\}_{\text{K80}}$
$\mathcal{B}_{\text{AT CG}}$	"	"	"	$\dots \cup \{(AG)X\}_{\text{K81}}$	$\dots \cup \{(AC)X\}_{\text{K80}}$
$\mathcal{B} \setminus \mathcal{B}_2$	"	"	"	$\dots \cup \{(AG)X\}_{\text{K81}}$	$\dots \cup \{(AC)X\}_{\text{K80}} \cup \{(AT)X\}_{\text{K80}}$

□

**Remark 23.** The sets of equations of Theorem 22 has been successfully used in [KDGC12] for model selection. Although the dimensions of these linear spaces are exponential in  $n$ , in practice it is not necessary to consider the full set of equations, but only those containing the patterns observed in the data. This is crucial for the applicability of the method, since the number of different columns in an alignment is really small compared to the dimension of these spaces.

**Example 24.** As an example, we compute a minimal system of equations for SSM, K81, K80, and JC69 in the case of 3 leaves.

*Equations for  $\mathcal{L}^{\text{SSM}}$ :*  $\mathbb{E}^{\text{SSM}}$  is composed of the following equations:

$$\begin{aligned}
 p_{AAA} &= p_{TTT}, & p_{AAC} &= p_{TTG}, & p_{AAG} &= p_{TTC}, & p_{AAT} &= p_{TTA}, \\
 p_{ACA} &= p_{TGT}, & p_{ACC} &= p_{TGG}, & p_{ACG} &= p_{TGC}, & p_{ACT} &= p_{TGA}, \\
 p_{AGA} &= p_{TCT}, & p_{AGC} &= p_{TCG}, & p_{AGG} &= p_{TCC}, & p_{AGT} &= p_{TCA}, \\
 p_{ATA} &= p_{TAT}, & p_{ATC} &= p_{TAG}, & p_{ATG} &= p_{TAC}, & p_{ATT} &= p_{TAA}, \\
 p_{CAA} &= p_{GTT}, & p_{CAC} &= p_{GTG}, & p_{CAG} &= p_{GTC}, & p_{CAT} &= p_{GTA}, \\
 p_{CCA} &= p_{GGT}, & p_{CCC} &= p_{GGG}, & p_{CCG} &= p_{GGC}, & p_{CCT} &= p_{GGA}, \\
 p_{CGA} &= p_{GCT}, & p_{CGC} &= p_{GCG}, & p_{CGG} &= p_{GCC}, & p_{CGT} &= p_{GCA}, \\
 p_{CTA} &= p_{GAT}, & p_{CTC} &= p_{GAG}, & p_{CTG} &= p_{GAC}, & p_{CTT} &= p_{GAA}.
 \end{aligned}$$

*Equations for  $\mathcal{L}^{\text{K81}}$ :*  $\mathbb{E}^{\text{K81}}$  is formed by  $\mathbb{E}^{\text{SSM}}$  and

$$\begin{aligned}
 p_{AAA} &= p_{CCC}, & p_{AAC} &= p_{CCA}, & p_{AAG} &= p_{CCT}, & p_{AAT} &= p_{CCG}, \\
 p_{ACA} &= p_{CAC}, & p_{ACC} &= p_{CAA}, & p_{ACG} &= p_{CAT}, & p_{ACT} &= p_{CAG}, \\
 p_{AGA} &= p_{CTC}, & p_{AGC} &= p_{CTA}, & p_{AGG} &= p_{CCT}, & p_{AGT} &= p_{CTG}, \\
 p_{ATA} &= p_{CGC}, & p_{ATC} &= p_{CGA}, & p_{ATG} &= p_{CGT}, & p_{ATT} &= p_{CGG}.
 \end{aligned}$$

*Equations for  $\mathcal{L}^{\text{K80}}$ :*  $\mathbb{E}^{\text{K80}}$  is formed by  $\mathbb{E}^{\text{K81}}$  and

$$\begin{aligned}
 p_{AAG} &= p_{GAA}, & p_{ACG} &= p_{GCA}, & p_{ACT} &= p_{GCT}, \\
 p_{AGA} &= p_{GAG}, & p_{AGC} &= p_{GAC}, & p_{AGG} &= p_{GAA}.
 \end{aligned}$$

Equations for  $\mathcal{L}^{\text{JC69}}$ :  $\mathbb{E}^{\text{JC69}}$  is formed by  $\mathbb{E}^{\text{K80}}$  and

$$p_{\text{AAC}} = p_{\text{ TTC}}, \quad p_{\text{ACA}} = p_{\text{ TCT}}, \quad p_{\text{ACC}} = p_{\text{ TCC}}, \quad p_{\text{ACG}} = p_{\text{ CAG}}, \quad p_{\text{ACG}} = p_{\text{ TCG}}.$$

**Identifiability of phylogenetic mixtures.** In this section we study the identifiability of phylogenetic mixtures. To this end, we use projective algebraic varieties and techniques from algebraic geometry. It is not our intention to give the reader a background on these tools, so we refer to the algebraic geometry book [Har92] and, more specifically, to [APRS11] for the usage of these techniques in the study of phylogenetic mixtures.

There is a natural isomorphism between the points lying in the hyperplane  $H$  considered above,  $H = \{p = (p_{\text{A...A}}, \dots, p_{\text{T...T}}) \in \mathcal{L} : \sum p_{x_1 \dots x_n} = 1\}$ , and the open affine subset  $\{p = [p_{\text{A...A}} : \dots : p_{\text{T...T}}] : \sum p_{x_1 \dots x_n} \neq 0\}$  of  $\mathbb{P}^{4^n - 1} = \mathbb{P}(\mathcal{L})$ . We use the notation  $[p_{\text{A...A}} : \dots : p_{\text{T...T}}]$  for projective coordinates (in contrast to  $(p_{\text{A...A}}, \dots, p_{\text{T...T}})$  used for affine coordinates). The *projective phylogenetic variety*  $\mathbb{P}V_T^{\mathcal{M}}$  associated to a phylogenetic tree  $T$  is the projective closure in  $\mathbb{P}(\mathcal{L})$  of the image of the stochastic parameterization  $\phi_T^{\mathcal{M}}$  defined above. That is, it is the smallest projective variety in  $\mathbb{P}(\mathcal{L})$  containing  $\text{Im } \phi_T^{\mathcal{M}}$  via the above isomorphism.

In what follows, we explain the relationship between this new variety and  $CV_T^{\mathcal{M}}$  and  $V_T^{\mathcal{M}}$ . By Remark 7, it becomes clear that  $CV_T^{\mathcal{M}}$  equals the affine cone over the projective phylogenetic variety  $\mathbb{P}V_T^{\mathcal{M}}$  (for the general Markov model, see also [AR08b, Proposition 1]). This implies that  $\dim CV_T^{\mathcal{M}} = \dim \mathbb{P}V_T^{\mathcal{M}} + 1$ , and if  $p = (p_{\text{A...A}}, \dots, p_{\text{T...T}})$  belongs to  $CV_T^{\mathcal{M}}$ , then  $q := [p_{\text{A...A}} : \dots : p_{\text{T...T}}]$  belongs to  $\mathbb{P}V_T^{\mathcal{M}}$ . Moreover, if  $\lambda := \sum p_{x_1 \dots x_n}$  is not zero, then  $(\frac{p_{\text{A...A}}}{\lambda}, \dots, \frac{p_{\text{T...T}}}{\lambda})$  is a point in the affine stochastic phylogenetic variety  $V_T^{\mathcal{M}}$ .

Before defining identifiability of mixtures, we consider the following construction of projective algebraic varieties.

**Definition 25.** Given two projective varieties  $X, Y \subset \mathbb{P}^m$ , the *join of  $X$  and  $Y$* ,  $X \vee Y$ , is the smallest variety in  $\mathbb{P}^m$  containing all lines  $\overline{xy}$  with  $x \in X$ ,  $y \in Y$ , and  $x \neq y$  (see [Har92, 8.1] for details). Similarly, one defines the *join of projective varieties*  $X_1, \dots, X_h \subset \mathbb{P}^m$ ,  $\vee_{i=1}^h X_i$ , as the smallest subvariety in  $\mathbb{P}^m$  containing all the linear varieties spanned by  $x_1, \dots, x_h$  with  $x_i \in X_i$  and  $x_i \neq x_j$ . It is known that

$$\dim(\vee_{i=1}^h X_i) \leq \min\left\{\sum_{i=1}^h \dim(X_i) + h - 1, m\right\}.$$

The right hand side of this inequality is usually known as the *expected dimension* of  $\vee_{i=1}^h X_i$ .

For instance, if we consider the join  $\bigvee_{i=1}^h \mathbb{P}V_{T_i}^{\mathcal{M}}$  for certain tree topologies  $T_i$  on the leaf set  $[n]$  and a given evolutionary model  $\mathcal{M}$ , then there is a (dominant rational) map

$$(7) \quad \mathbb{P}V_{T_1}^{\mathcal{M}} \times \mathbb{P}V_{T_2}^{\mathcal{M}} \times \dots \times \mathbb{P}V_{T_h}^{\mathcal{M}} \times \mathbb{P}^{h-1} \dashrightarrow \bigvee_{i=1}^h \mathbb{P}V_{T_i}^{\mathcal{M}} \subset \mathbb{P}(\mathcal{L}),$$

which is the projective closure of the parameterization  $\phi_{T_1} \vee \dots \vee \phi_{T_h}$  defined by

$$\begin{aligned} \text{Par}_{s\mathcal{M}}(T_1) \times \dots \times \text{Par}_{s\mathcal{M}}(T_h) \times \Omega &\longrightarrow \mathcal{L} \\ (\xi_1, \dots, \xi_h, \mathbf{a}) &\mapsto \sum_j a_j \phi_{T_i}^{\mathcal{M}}(\xi_i). \end{aligned}$$

Here,  $\Omega = \{\mathbf{a} = (a_1, \dots, a_h) \mid \sum_i a_i = 1\}$  is isomorphic to an affine open subset of  $\mathbb{P}^{h-1}$ . In this setting, an  $h$ -mixture on  $\{T_1, \dots, T_h\}$  corresponds to a point in the variety  $\bigvee_{i=1}^h \mathbb{P}V_{T_i}^{\mathcal{M}}$ . We will use this algebraic variety to study the identifiability of phylogenetic mixtures.

When considering unmixed models  $\mathcal{M}$  on trivalent trees on  $n$  taxa, *generic identifiability of the tree topology* is equivalent to the projective varieties  $\mathbb{P}V_T^{\mathcal{M}}$  and  $\mathbb{P}V_{T'}^{\mathcal{M}}$  being different when  $T \neq T'$  (see [AR08a]). The identifiability of the continuous parameters must take into account the possibility of permuting the labels of the states at the interior nodes, as such permutations give rise to the same joint distribution at the leaves. In the language of algebraic geometry, *generic identifiability of the continuous parameters* of the model implies that the map  $\phi_T^{\mathcal{M}}$  is generically finite (i.e. the preimage of a generic point is a finite number of points; see [AR08a]). In this case, the fiber dimension Theorem [Har92, Theorem 11.12] applies and we have that  $\dim \mathbb{P}V_T^{\mathcal{M}}$  is equal to the number of stochastic parameters of the model,  $\dim \text{Par}_{s\mathcal{M}}(T)$ . Therefore, if the continuous parameters are generically identifiable for the unmixed trees under  $\mathcal{M}$ , then the dimension of the variety  $\mathbb{P}V_T^{\mathcal{M}}$  is the same for all trivalent tree topologies on  $n$  taxa. This dimension is denoted by  $s_{\mathcal{M}}(n)$ .

**Example 26.** The tree topologies and the continuous parameters are generically identifiable for the unmixed equivariant models JC69, K80, K81, SSM, and GMM on trees with any number of leaves (see [AR06a] and [CFS11, Corollary 3.9]).

From now on we only consider trees without nodes of degree 2, so that the number of free stochastic parameters on a phylogenetic tree on  $n$  taxa under  $\mathcal{M}$  is  $\leq s_{\mathcal{M}}(n)$ .

We recall the definition of generic identifiability of the tree topologies on  $h$ -mixtures (see [APRS11]).

**Definition 27.** The *tree topologies* on  $h$ -mixtures under  $\mathcal{M}$  are *generically identifiable* if for any set of trivalent tree topologies  $\{T_1, \dots, T_h\}$  and a generic choice of  $(\xi_1, \dots, \xi_h, \mathbf{a}) \in \text{Par}_{s\mathcal{M}}(T_1) \times \dots \times \text{Par}_{s\mathcal{M}}(T_h) \times \Omega$ , the equality

$$\phi_{T_1} \vee \dots \vee \phi_{T_h}(\xi_1, \dots, \xi_h, \mathbf{a}) = \phi_{T'_1} \vee \dots \vee \phi_{T'_h}(\xi'_1, \dots, \xi'_h, \mathbf{a}'),$$

for tree topologies  $\{T'_1, \dots, T'_h\}$  and  $(\xi'_1, \dots, \xi'_h, \mathbf{a}') \in \text{Par}_{s_{\mathcal{M}}}(T'_1) \times \dots \times \text{Par}_{s_{\mathcal{M}}}(T'_h) \times \Omega$  implies

$$\{T_1, \dots, T_h\} = \{T'_1, \dots, T'_h\}.$$

In terms of algebraic varieties this is equivalent to saying that the variety  $\bigvee_{i=1}^h \mathbb{P}V_{T'_i}^{\mathcal{M}}$  is not contained in  $\bigvee_{i=1}^h \mathbb{P}V_{T_i}^{\mathcal{M}}$  and vice versa.

The tree topologies are the discrete parameters of  $h$ -mixtures. When considering the continuous parameters of  $h$ -mixtures, the above mentioned label-swapping can be disregarded. We give the following definition according to [RS12].

**Definition 28.** The *continuous parameters* of  $h$ -mixtures on  $T_1, \dots, T_h$  under an evolutionary model  $\mathcal{M}$  are *generically identifiable* if, for a generic choice of stochastic parameters  $(\xi_1, \dots, \xi_h, \mathbf{a})$ , the equality

$$\phi_{T_1} \vee \dots \vee \phi_{T_h}(\xi_1, \dots, \xi_h, \mathbf{a}) = \phi_{T_1} \vee \dots \vee \phi_{T_h}(\xi'_1, \dots, \xi'_h, \mathbf{a}')$$

for stochastic parameters  $(\xi'_1, \dots, \xi'_h, \mathbf{a}')$  implies that there is a permutation  $\sigma \in \mathfrak{S}_h$  such that  $\sigma \cdot (T_1, \dots, T_h) = (T_1, \dots, T_h)$ ,  $\xi'_i = \xi_{\sigma(i)}$ , and  $a'_i = a_{\sigma(i)}$  for  $i = 1, \dots, h$ . In other words, we only allow swapping of the continuous parameters when at least two tree topologies coincide.

**Definition 29.** An  $h$ -mixture under a model  $\mathcal{M}$  is said to be *identifiable* if both its tree topologies and its continuous parameters are generically identifiable.

In terms of algebraic varieties, generic identifiability of continuous parameters on  $h$ -mixtures implies that the generic fibers (i.e. preimages of generic points) of the map  $\phi_{T_1} \vee \dots \vee \phi_{T_h}$  are finite. In this case, the fiber dimension theorem applied to (7) (cf. [Har92, Theorem 11.12]) gives

$$\dim(\bigvee_{i=1}^h \mathbb{P}V_{T_i}) = \sum_{i=1}^h \dim(\mathbb{P}V_{T_i}) + h - 1.$$

The following result demonstrates the need for careful inspection of identifiability of mixtures with many components (i.e. large values of  $h$ ).

**Theorem 30.** *Let  $[n]$  be a set of taxa and  $\mathcal{M}$  be an evolutionary model for which the continuous parameters are generically identifiable on trivalent (unmixed) trees. In addition, let  $s_{\mathcal{M}}(n)$  be the dimension of  $\mathbb{P}V_T^{\mathcal{M}}$  for any trivalent tree  $T$ , and set  $h_0(n) := \frac{\dim \mathcal{D}_{\mathcal{M}}}{s_{\mathcal{M}}(n)+1}$ . Then the  $h$ -mixtures of trees on  $[n]$  evolving under  $\mathcal{M}$  are not identifiable for  $h \geq h_0(n)$ .*

**Remark 31.** Note that, in the above definition of  $h_0(n)$ ,  $\dim \mathcal{D}_{\mathcal{M}}$  also depends on  $n$ .

**Corollary 32.** *Let  $[n]$  be a set of taxa and  $\mathcal{M}$  be one of the equivariant models JC69, K80, K81, SSM, or GMM. Then the phylogenetic  $h$ -mixtures under these models are not identifiable for  $h \geq h_0(n)$ , where*

$$h_0(n) = \begin{cases} \frac{4^n}{12(2n-3)+4}, & \text{if } \mathcal{M} = \text{GMM}, \\ \frac{2^{2n-1}}{6(2n-3)+2}, & \text{if } \mathcal{M} = \text{SSM}, \\ \frac{4^{n-1}}{3(2n-3)+1}, & \text{if } \mathcal{M} = \text{K81}, \\ \frac{2^{2n-3}+2^{n-2}}{2(2n-3)+1}, & \text{if } \mathcal{M} = \text{K80}, \\ \frac{2^{2n-3}+3 \cdot 2^{n-2}+1}{3(2n-2)}, & \text{if } \mathcal{M} = \text{JC69}. \end{cases}$$

*Proof.* Theorem 17 shows that  $\mathcal{L}^{\mathcal{M}} = \mathcal{D}_{\mathcal{M}}$  and Proposition 20 gives the dimension of  $\mathcal{L}^{\mathcal{M}}$  in each case. Next, we calculate:  $s_{\text{GMM}}(n) = 12(2n-3) + 3$ ,  $s_{\text{SSM}}(n) = 6(2n-3) + 1$ ,  $s_{\text{K81}}(n) = 3(2n-3)$ ,  $s_{\text{K80}}(n) = 2(2n-3)$ , and  $s_{\text{JC69}}(n) = 2n-3$ . Applying Theorem 30, we conclude the proof.  $\square$

**Example 33.** Consider the Kimura 3-parameter model K81 on  $n = 4$  taxa. For any  $h \geq 4$ , phylogenetic  $h$ -mixtures are not identifiable by Corollary 32. We are not aware of any result proving that mixtures of 2 or 3 different tree topologies under this model are identifiable (either for the tree parameters or for the continuous parameters).

**Example 34.** Consider the Jukes-Cantor model JC69 on  $n = 4$  taxa. Then Corollary 32 tells us that for  $h \geq 3$ ,  $h$ -mixtures are not identifiable. Therefore, for this particular model on four taxa the cases in which the identifiability holds are known: the tree and the continuous parameters are generically identifiable for the unmixed model; the tree parameters are generically identifiable for 2-mixtures [APRS11, Theorem 10]; the continuous parameters are generically identifiable for 2-mixtures on different tree topologies and not identifiable for the same tree topology [APRS11, Theorem 23]; neither the continuous parameters nor the tree topologies are generically identifiable for mixtures with more than two components (Corollary 32).

*Proof of Theorem 30.* Let  $\text{edim}(h) := hs_{\mathcal{M}}(n) + h - 1$ . Then the variety  $\bigvee_{i=1}^h \mathbb{P}V_{T_i}$  has dimension  $\leq \text{edim}(h)$ . Indeed, as  $\bigvee_i \phi_{T_i}$  is a parameterization of an open subset of  $\bigvee_{i=1}^h \mathbb{P}V_{T_i}$ , then the dimension of  $\bigvee_{i=1}^h \mathbb{P}V_{T_i}$  is less than or equal to  $\sum \dim \mathbb{P}V_{T_i} + h - 1$ . Moreover, the dimension of  $\mathbb{P}V_{T_i}$  is equal to  $s_{\mathcal{M}}(n)$  if  $T_i$

is trivalent (since the continuous parameters for the unmixed models under consideration are generically identifiable) and is less than  $s_{\mathcal{M}}(n)$  for non-trivalent trees. Therefore,  $\dim(\vee_{i=1}^h \mathbb{P}V_{T_i}) \leq \text{edim}(h)$ .

If all  $T_i$  are trivalent trees, then  $\sum \dim \mathbb{P}V_{T_i} + h - 1 = \text{edim}(h)$  and, therefore,  $\dim(\vee_{i=1}^h \mathbb{P}V_{T_i}) < \text{edim}(h)$  if and only if  $\dim(\vee_{i=1}^h \mathbb{P}V_{T_i}) < \sum \dim \mathbb{P}V_{T_i} + h - 1$ . Moreover, by the fiber dimension theorem applied to  $\vee \phi_{T_i}$ , the equality of dimensions holds if and only if the generic fiber of  $\vee \phi_{T_i}$  has dimension 0. In particular, if  $\dim(\vee_{i=1}^h \mathbb{P}V_{T_i}) < \text{edim}(h)$ , then the continuous parameters of this phylogenetic mixture are not identifiable.

As  $h_0(n) = \frac{\dim \mathcal{D}_{\mathcal{M}}}{s_{\mathcal{M}}(n)+1}$ , we have that  $\text{edim}(h_0(n)) = h_0(n)(s_{\mathcal{M}}(n) + 1) - 1 = \dim \mathcal{D}_{\mathcal{M}} - 1$ . Now, we fix an  $h \in \mathbb{N}$  with  $h \geq h_0(n)$ , so that  $\text{edim}(h) \geq \dim(\mathcal{D}_{\mathcal{M}}) - 1$ .

There are two possible scenarios:

(a) For any set of tree topologies  $\{T_1, \dots, T_h\}$ , the dimension of  $\vee_{i=1}^h \mathbb{P}V_{T_i}$  is less than  $\dim(\mathcal{D}_{\mathcal{M}}) - 1$ .

(b) There exists a set of tree topologies  $\{T_1, \dots, T_h\}$  for which  $\dim(\vee_{i=1}^h \mathbb{P}V_{T_i}) = \dim(\mathcal{D}_{\mathcal{M}}) - 1$ .

Case (a) implies that the dimension of  $\vee_{i=1}^h \mathbb{P}V_{T_i}$  is less than  $\text{edim}(h)$  for any set of trivalent tree topologies  $\{T_1, \dots, T_h\}$ . Based on the conclusions drawn above, this implies that the continuous parameters are not generically identifiable.

In case (b),  $\vee_{i=1}^h \mathbb{P}V_{T_i}$  coincides with  $\mathbb{P}(\mathcal{D}_{\mathcal{M}})$ . Indeed,  $\vee_{i=1}^h \mathbb{P}V_{T_i}$  is contained in  $\mathbb{P}(\mathcal{D}_{\mathcal{M}})$ , both varieties are irreducible, and  $\dim(\vee_{i=1}^h \mathbb{P}V_{T_i}) = \dim(\mathcal{D}_{\mathcal{M}}) - 1 = \dim(\mathbb{P}(\mathcal{D}_{\mathcal{M}}))$ , which implies that both varieties coincide. In particular, any  $h$ -mixture (which is a point in  $\mathbb{P}(\mathcal{D}_{\mathcal{M}})$ ) would be contained in  $\vee_{i=1}^h \mathbb{P}V_{T_i}$ , and therefore the topologies are not generically identifiable.  $\square$

**Remark 35.** The negative result of Theorem 30 should be complemented with the following positive result of Rhodes and Sullivant in [RS12]: if  $\mathcal{M} = \text{GMM}$  and one restricts to  $h$ -mixtures on the same trivalent tree topology  $T$ , then the tree topology and the continuous parameters are generically identifiable if  $h < 4^{\lceil \frac{n}{4} \rceil - 1}$ .

#### AUTHOR'S CONTRIBUTIONS

All authors contributed equally and the author names order is alphabetical.



## ACKNOWLEDGEMENTS

All authors are partially supported by Generalitat de Catalunya, 2009 SGR 1284. Research of the first and second authors partially supported by Ministerio de Educación y Ciencia MTM2009-14163-C02-02. We would like to thank both referees for very useful comments that led to major improvements.

## REFERENCES

- [APRS11] E S Allman, S Petrovic, J A Rhodes, and S Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 8:710–720, 2011.
- [AR03] ES Allman and JA Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 186(2):113–144, 2003.
- [AR04] ES Allman and JA Rhodes. Quartets and parameter recovery for the general Markov model of sequence mutation. *Applied Mathematics Research Express*, 2004(4):107–131, 2004.
- [AR06a] ES Allman and JA Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *Journal of Computational Biology*, 13:1101–1113, 2006.
- [AR06b] ES Allman and JA Rhodes. Phylogenetic invariants for stationary base composition. *Journal of Symbolic Computation*, 41(2):138 – 150, 2006.
- [AR08a] ES Allman and JA Rhodes. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. *Mathematical Biosciences*, 211:18–33, 2008.
- [AR08b] ES Allman and JA Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Advances in Applied Mathematics*, 40:127–148, 2008.
- [CF87] J Cavender and J Felsenstein. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71, 1987.
- [CFS11] M Casanellas and J Fernandez-Sanchez. Relevant phylogenetic invariants of evolutionary models. *Journal de Mathématiques Pures et Appliquées*, 96:207–229, 2011.
- [CH11] J Chai and E A Housworth. On Rogers’s proof of identifiability for the GTR + Gamma + I model. *Systematic Biology*, 2011.
- [Cha96] J T Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences*, 137(1):51–73, 1996.
- [CS05] M Casanellas and S Sullivant. The strand symmetric model. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for computational biology*, chapter 16. Cambridge University Press, 2005.
- [DK09] J Draisma and J Kuttler. On the ideals of equivariants tree models. *Mathematische Annalen*, 344:619–644, 2009.
- [Fel03] J Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., 2003.
- [FL92] Y-X Fu and WH Li. Construction of linear invariants in phylogenetic inference. *Mathematical Biosciences*, 109(2):201 – 228, 1992.
- [Har92] J Harris. *Algebraic geometry. A first course*, volume 133 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1992.
- [JC69] T H Jukes and C R Cantor. Evolution of protein molecules. In H N Munro, editor, *Mammalian Protein Metabolism*, volume 3, pages 21–132. Academic Press, 1969.

- [KDGC12] A Kedzierska, M Drton, R Guigó, and M Casanellas. SPI<sub>n</sub>: model selection for phylogenetic mixtures via linear invariants. *Molecular Biology and Evolution*, 29:929–937, 2012.
- [Kim80] M Kimura. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.
- [Kim81] M Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences*, 78:454–458, 1981.
- [Lak87] JA Lake. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, 4:167–191, 1987.
- [MMS08] FA Matsen, E Mossen, and M Steel. Mixed-up trees: The structure of phylogenetic mixtures. *Bulletin of Mathematical Biology*, 70:1115–1139, 2008.
- [Pos01] D Posada. The effect of branch length variation on the selection of models of molecular evolution. *Journal of Molecular Evolution*, 52:434–444, 2001.
- [RS12] JA Rhodes and S Sullivant. Identifiability of large phylogenetic mixture models. *Bulletin of Mathematical Biology*, 74:212–231, 2012.
- [Ser77] JP Serre. *Linear representations of finite groups*. Springer-Verlag, New York, 1977. Translated from the second French edition by Leonard L. Scott, Graduate Texts in Mathematics, Vol. 42.
- [SFSJ12] J Sumner, J Fernández-Sánchez, and P Jarvis. On lie markov models. *Journal of Theoretical Biology*, 298:16–31, 2012.
- [SHP98] MA Steel, MD Hendy, and D Penny. Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results. *Discrete Applied Mathematics*, 88(1-3):367–396, 1998.
- [SHSE92] MA Steel, MD Hendy, LA Székely, and PL Erdős. Spectral analysis and a closest tree method for genetic sequences. *Applied Mathematics Letters. An International Journal of Rapid Publication*, 5:63–67, 1992.
- [SS03] C Semple and M Steel. *Phylogenetics*, volume 24 of Oxford Lecture Series in Mathematics and its Applications. Oxford University Press, Oxford, 2003.
- [SS05] B Sturmfels and S Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12:204–228, 2005.
- [SV07] D Stefanovic and E Vigoda. Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *Journal of Computational Biology*, 14:156–189, 2007.

DEPARTAMENT DE MATEMÀTICA APLICADA I. ETSEIB. UNIVERSITAT POLITÈCNICA DE CATALUNYA. AVINGUDA DIAGONAL 647. 08028 BARCELONA. SPAIN.

*E-mail address:* `marta.casanellas@upc.edu`

DEPARTAMENT DE MATEMÀTICA APLICADA I. ETSEIB. UNIVERSITAT POLITÈCNICA DE CATALUNYA. AVINGUDA DIAGONAL 647. 08028 BARCELONA. SPAIN.

CENTRE FOR GENOMIC REGULATION (CRG). DR. AIGUADER 88. 08003 BARCELONA. SPAIN.