

THE SPACE OF PHYLOGENETIC MIXTURES FOR EQUIVARIANT MODELS

MARTA CASANELLAS, JESÚS FERNÁNDEZ-SÁNCHEZ, AND ANNA KEDZIERSKA

ABSTRACT. The selection of the most suitable evolutionary model to analyze the given molecular data is usually left to biologist's choice. In his famous 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 open the question regarding whether these equations were enough for characterizing the evolutionary model.

Here we prove that, for equivariant models of evolution, the space of distributions satisfying these linear equations coincides with the space of distributions arising from mixtures of trees on a set of taxa. In other words, we prove that an alignment is produced from a mixture of phylogenetic trees under an equivariant evolutionary model if and only if its distribution of column patterns satisfies the linear equations mentioned above. Moreover, for each equivariant model and for any number of taxa, we provide a set of linearly independent equations defining this space of phylogenetic mixtures. This is a powerful tool that has already been successfully used in model selection. We also use the results obtained to study identifiability issues for phylogenetic mixtures.

1. INTRODUCTION

In phylogenetics, the goal is to reconstruct the ancestral relationships among organisms. Most of the widely used phylogenetic reconstruction methods are based on mathematical models describing the molecular evolution of DNA. The problem of choosing the most suitable model for the given data is usually left to biologist choice, and there are no fully reliable methods to choose the best model (cf. [Pos01]).

In this paper we address the following question: in accordance to Darwin's theory that evolution occurs following a tree, how can the data coming from a particular evolutionary model be characterized? In other words, are there any invariants of the DNA patterns that have evolved following a tree (or a mixture

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.

of trees, as we will see below) under a particular model? The answer to these questions leads to a complete characterization of the evolutionary model and therefore can be used as a selection criterion for the most suitable evolutionary model for the given data.

Here we explain our motivation to solve this problem. It is well known that the expected probabilities of nucleotides observed at the leaves of a phylogenetic tree satisfy a collection of equalities if the tree evolves under certain models (see for instance [Fel03, p.375]). It was already pointed out by [FL92], [SHSE92] or [Fel03], that these equalities (referred to as *linear invariants*) could potentially be used to test the model of base change; but, how could it be guaranteed that there were no more equalities to be used?

We answer the questions above for equivariant models of molecular evolution ([DK09], [CFS11]). These are Markov processes on trees whose transition matrices satisfy certain symmetries and include Jukes-Cantor model, Kimura 2 and 3 parameters, strand symmetric model and general Markov model. Our main result in section 4 states that, if evolution occurs according to trees (or even mixtures of trees), then the model of evolution is determined by a linear space. By exhaustively studying the group of symmetries of these models, we also give an easy and combinatorial way of determining the equations of this linear space. In the paper [KDGC11] these equations are successfully used for model selection in phylogenetics.

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} ; that is, the set of points that are a linear combination of points lying in the phylogenetic varieties $CV_T^{\mathcal{M}}$ (see section 2 for an explanation on this terminology). In biological words, this is the set of alignments whose columns have evolved following the model \mathcal{M} under a phylogenetic tree (not necessarily the same tree in the whole alignment, and not necessarily the same transition matrices). In phylogenetics, the hypothesis that the sites of an alignment are independent and identically distributed is often used in the most simple models. When one removes the assumption “identically distributed” and replaces it by “distributed according to the same evolutionary model” then one obtains a phylogenetic mixture. This phenomena is needed to explain heterogeneous evolutionary processes when data comprises multiple genes or selected codon positions. Among a plethora of applications, phylogenetic mixtures are used in the orthology prediction, gene and genome annotation, species tree reconstruction or drug target identification. In the usual setting, phylogenetic mixtures are modeled by assuming rate variation across sites (see [SS03] for more information on these concepts).

As a byproduct we are able to determine the dimension of these linear spaces and use it to give an upper bound on the number of mixtures that can be used in a phylogenetic reconstruction problem. This is part of the *identifiability* issue in phylogenetic mixtures: determine conditions that guarantee that the model parameters (trees and continuous parameters) can be recovered from the data. This problem is crucial for justifying methods as maximum likelihood and has been extensively studied recently, but only few results are known (see for instance [AR06a], [APRS10], [SV07], [RS],[CH11]). We explain this topic in detail in section 5.

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

2. PRELIMINARIES

Let n a number and denote by $[n]$ the set $\{1, 2, \dots, n\}$. For biological purposes, we think of $[n]$ as set of DNA sequences associated to certain taxa and we consider trees as connected acyclic graphs whose n leaves are bijectively labelled by the set $[n]$. Let \mathbb{T}_n be the set of tree topologies (up to isomorphism) whose leaves are labelled by $[n]$. Trees in \mathbb{T}_n are allowed to have any degree in its internal vertices. When the internal vertices of a tree $T \in \mathbb{T}_n$ have degree 3, we say that the tree is *trivalent*.

Here we introduce the definitions needed in the sequel.

We fix an ordered set $B = \{b_1, b_2, \dots, b_k\}$ and we think of it as a basis of a vector space $W := \langle B \rangle_{\mathbb{C}}$. For example, for most applications we use $B = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ and think of its elements as nucleotides in a DNA sequence.

Definition 2.1. A *phylogenetic tree on W* is a tree T that has the vector space $W_v := W$ associated to each vertex v of T . Usually the same notation T is used to represent both the graph and the phylogenetic tree. Elements of B at the vertices of T are thought as states of discrete random variables of the vertices.

Definition 2.2. Let T be a phylogenetic tree on W and assume that a distinguished vertex r of T (usually called the *root*) is given, inducing therefore an orientation in all edges of T . If e is an edge of T , we write e_0 and e_1 for the origin and final vertices of the edge e , respectively. An *evolutionary presentation of a phylogenetic tree T* is a vector $\pi = (\pi_{b_1}, \pi_{b_2}, \dots, \pi_{b_k}) \in W_r$, together with a collection of maps $\mathbf{A} = (A^{e_0, e_1})_{e \in E(T)}$ where each A^{e_0, e_1} belongs to $\text{Hom}(W_{e_0}, W_{e_1})$.

From now on, we will identify vectors in W with its coordinates in the basis B written as a column vector. Similarly, we will identify the set $\text{Hom}(W, W)$ with

the set of matrices with k rows and k columns and entries in the complex field by mapping any linear map to its matrix in the basis B . We take the convention that the matrices $A = A^{e_0, e_1}$ in an evolutionary presentation act on W from the right (i.e. the action is $\omega^t \in W_{e_0} \mapsto \omega^t A \in W_{e_1}$).

From now on, the vector $(1, 1, \dots, 1) \in W$ will be denoted by $\mathbf{1}$.

Definition 2.3. 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 every $\pi \neq 0$ in W_0 , together with a multiplicatively closed (grupoid) subspace Mod of $\text{Hom}(W, W)$. We will usually denote it by $\mathcal{M} = (W_0, Mod)$.

If T is a rooted phylogenetic tree on W , then T *evolves under the algebraic evolutionary model* \mathcal{M} if its evolutionary presentations lie in Mod and the vector π at the root belongs to W_0 . We denote by $\text{Par}_{\mathcal{M}}(T) = W_0 \times \left(\prod_{e \in E(T)} Mod \right)$.

Remark 2.4. The condition $\mathbf{1}^T \pi \neq 0$ for every $\pi \in W_0$ in the definition above means that the sum of the coordinates of the vectors in W_0 is different from zero, unless the vector is already 0. Since the vectors in W_0 represent the possible distributions for the root in the tree T , this condition is biologically meaningful and implies no significant restriction for our evolutionary models.

Below we give some well known examples of evolutionary models.

Definition 2.5. 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* is defined by taking Mod equal to $\text{Hom}_G(W, W)$, that is,

$$\text{Hom}_G(W, W) = \{A \in M_{k,k}(\mathbb{C}) \mid AP_g = P_g A, \forall g \in G\}$$

and $W_0 = \{\pi \in W \mid P_g \pi = \pi \forall g \in G\}$. It is clear that the above subsets define vector subspaces of $\text{Hom}(W, W)$ and W_0 . On the other hand, if $A_1, A_2 \in \text{Hom}_G(W, W)$, then

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

so $A_1 A_2 \in \text{Hom}_G(W, W)$. Therefore, equivariant models provide a wide family of examples of algebraic evolutionary models in the sense of Definition 2.3. For example, if $B = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ and

- $G = \mathfrak{S}_4$, this is the *algebraic Jukes-Cantor model* JC69,
- $G = \langle (\mathbf{ACGT}), (\mathbf{AG}) \rangle$, this is the *algebraic Kimura 2-parameter model* K80,
- $G = \langle (\mathbf{AC})(\mathbf{GT}), (\mathbf{AG})(\mathbf{CT}) \rangle$, this is the *algebraic Kimura 3-parameter model* K81,

- $G = \langle (\text{AT})(\text{CG}) \rangle$, it is known as the *strand symmetric model SSM*, and
- $G = \langle id \rangle$, this is the *general Markov model GMM*.

For an equivariant model \mathcal{M} we denote by $G_{\mathcal{M}}$ the corresponding group.

Notation 2.6. If $\omega = (\omega_1, \dots, \omega_k)$ is a vector in W , then D_ω denotes the square matrix with ω on the diagonal and zeros elsewhere.

Example 2.7. ([AR06b]) Let $\pi \in W$ be such that $\pi^t \mathbf{1} \neq 0$. The π -stable base distribution model (π -SBD) is given by taking $W_0 = \langle \pi \rangle \subset W$ and letting Mod be the set matrices for which π is a left eigenvector of every $A \in Mod$. Notice that the π -SBD satisfy the conditions of algebraic evolutionary model of Definition 2.3: Mod is a vector space and it is multiplicatively closed. By adding the conditions that π has eigenvalue 1 for every $A \in Mod$ and the entries of each row of each matrix A sum to one, we obtain the stable base distribution model defined in [AR06b] (see Definition 2.14).

For example, JC69, K80 and K81 are examples of π -SBD submodels with $\pi = (1, 1, 1, 1)$. The model SSM is not an SBD model since matrices in Mod do not share a fixed left eigenvector.

Definition 2.8. Given a phylogenetic tree T on W , $T \in \mathbb{T}_n$, an $[n]$ -tensor is any element of

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

Notation 2.9. We will denote by $\mathcal{B} = B^n$ the set of n -words in B ,

$$\mathcal{B} = \{X = (\mathbf{x}_1, \dots, \mathbf{x}_n) : \mathbf{x}_i \in B\}.$$

For the sake of simplicity in our notation, sometimes it will be convenient to identify every word $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ with the tensor $\mathbf{x}_1 \otimes \dots \otimes \mathbf{x}_n \in \mathcal{L}$ and consequently, we will identify \mathcal{B} with the natural basis of \mathcal{L} . Notice that a distribution $p = (p_{b_1 \dots b_1}, \dots, p_{b_k \dots b_k})$ on the set of patterns in B at the leaves of a tree can be viewed as the tensor in \mathcal{L} having these coordinates in the basis \mathcal{B} , that is

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

Definition 2.10. Given an algebraic evolutionary model \mathcal{M} , the *parametrization* of a rooted phylogenetic tree T on W evolving under the model \mathcal{M} is the map

$$\Psi_T^{\mathcal{M}} : \text{Par}_{\mathcal{M}}(T) \longrightarrow \mathcal{L} = \otimes_{[n]} W$$

that correspond to a hidden Markov process on the tree T when we restrict to stochastic matrices and distributions in W_0 (leaves correspond to observed random variables and the interior nodes to hidden variables). That is, if the tree

is rooted and directed from the root r , then the parametrization of T is the map

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

where

$$(1) \quad p_{\mathbf{x}_1 \dots \mathbf{x}_n} = \sum_{\mathbf{x}_v \in B, v \in \text{Int}(T)} \pi_{\mathbf{x}_r} \prod_{v \in N(T) \setminus \{r\}} A_{\mathbf{x}_{pa(v)}, \mathbf{x}_v}^{e_{pa(v), e_v}},$$

\mathbf{x}_u denotes the state at the vertex u , $pa(v)$ is the parent node of v , $\pi = (\pi_{\mathbf{x}})_{\mathbf{x} \in B}$ are the coordinates in the basis B of a vector associated to the root, and if $v = i$ is a leaf node, then $\mathbf{x}_v = \mathbf{x}_i$.

Note that the position of the root plays a role in the above parameterization. However, the following lemma shows that under some assumptions, its image is independent of it. Let T_u be a tree rooted at a vertex u , and consider a vertex v adjacent to u . Then, rooting T_u at v induces the opposite orientation on the edge \bar{e} delimited by u and v . Write T_v for the new oriented rooted tree. Then we have the following lemma whose proof is left to the reader.

Lemma 2.11. *Let T_u be evolving under an algebraic evolutionary model $\mathcal{M} = (W_0, \text{Mod})$. Let (π, \mathbf{A}) be an evolutionary presentation on T_u such that π has all its entries different from 0 and let $\tilde{\pi}^t = \pi^t A^{\bar{e}}$. Assume also that all the entries of $\tilde{\pi} \in W_0$ are different from 0 and $D_{\tilde{\pi}}^{-1}(A^{\bar{e}})^t D_{\pi}$ belongs to Mod . Then, letting*

$$\tilde{A}^e := \begin{cases} D_{\tilde{\pi}}^{-1}(A^{\bar{e}})^t D_{\pi}, & \text{if } e = \bar{e}, \\ A^e, & \text{otherwise} \end{cases}$$

and $\tilde{\mathbf{A}} = (\tilde{A}^e)_{e \in E(T_v)}$, we have $\Psi_{T_u}^{\mathcal{M}}(\pi, \mathbf{A}) = \Psi_{T_v}^{\mathcal{M}}(\tilde{\pi}, \tilde{\mathbf{A}})$.

Therefore, for the models satisfying the conditions of the previous lemma the image of the map $\Psi_T^{\mathcal{M}}$ does not depend on the position of the root. We call these models root-independent models:

Definition 2.12. We say that an algebraic evolutionary model $\mathcal{M} = (W_0, \text{Mod})$ is *root-independent* if it satisfies

- (i) $\tilde{\pi}^t := \pi^t A$ belongs to W_0 for all $\pi \in W_0$ and all $A \in \text{Mod}$, and
- (ii) $D_{\tilde{\pi}}^{-1} A^t D_{\pi} \in \text{Mod}$ whenever $D_{\tilde{\pi}}^{-1}$ does exist.

Example 2.13. Equivariant models as well as the stable base distribution models are root-independent (we let the reader check it for equivariant models and we refer to [AR06b] for the SBD model).

Definition 2.14. A *stochastic evolutionary model* $s\mathcal{M}$ is specified by a subset sW_0 of vectors in W whose entries sum to one, together with a multiplicatively closed set $s\text{Mod}$ of complex matrices whose rows sum to one.

Notice that the rows of a matrix sum to one if and only if $\mathbf{1}$ is a right-eigenvector corresponding to eigenvalue 1. Hence, for a stochastic evolutionary model $s\mathcal{M}$, the space of matrices $sMod$ is not a vector subspace anymore.

Example 2.15. If $\mathcal{M} = (W_0, Mod)$ is an algebraic evolutionary model, define $s\mathcal{M} = (sW_0, sMod)$ as $sW_0 = \{\pi \in W_0 : \mathbf{1}^t \pi = 1\}$ and $sMod = \{A \in Mod : A\mathbf{1} = \mathbf{1}\}$. Then, $s\mathcal{M}$ is a stochastic evolutionary model.

Definition 2.16. The parameterization $\Psi_T^{\mathcal{M}}$ of a rooted tree T evolving under a model \mathcal{M} restricts to a polynomial map $\phi_T^{\mathcal{M}}$ from

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

to the hyperplane $H \subset \mathcal{L}$ defined by

$$H = \left\{ p \in \mathcal{L} : \sum_{\mathbf{x}_1, \dots, \mathbf{x}_n \in B} p_{\mathbf{x}_1 \dots \mathbf{x}_n} = 1 \right\}.$$

From now on, we will refer to this map as the *stochastic parametrization*.

The map $\Psi_T^{\mathcal{M}}$ restricted to distributions in W_0 and stochastic matrices in Mod assigns to each set of parameters the corresponding distribution of patterns in B at the leaves of the tree and therefore its image lies on the standard simplex in $\mathcal{L} = \otimes_{[n]} W$. This justifies why the image of the stochastic parametrization $\phi_T^{\mathcal{M}}$ lies in H .

We proceed to define algebraic varieties associated to the parameterization maps. For background in algebraic geometry see [Har92].

Definition 2.17. The *affine phylogenetic variety* $CV_T^{\mathcal{M}}$ associated to a phylogenetic tree T on W is

$$CV_T^{\mathcal{M}} := \overline{\{\Psi_T^{\mathcal{M}}(\pi_r, \mathbf{A}) : (\pi_r, \mathbf{A}) \in \text{Par}_{\mathcal{M}}(T)\}}$$

where the closure is taken in the Zariski topology. Equivalently, $CV_T^{\mathcal{M}}$ is the smallest algebraic set containing the image of $\Psi_T^{\mathcal{M}}$.

The *affine stochastic phylogenetic variety* $V_T^{\mathcal{M}}$ associated to a phylogenetic tree T on W is

$$V_T^{\mathcal{M}} := \overline{\{\phi_T^{\mathcal{M}}(\pi_r, \mathbf{A}) : (\pi_r, \mathbf{A}) \in \text{Par}_{s\mathcal{M}}(T)\}} \subset H$$

where the closure is taken in the Zariski topology.

There is a natural isomorphism between the points lying in the hyperplane $H = \{p = (p_{b_1 \dots b_1}, \dots, p_{b_k \dots b_k}) \in \mathcal{L} : \sum p_{\mathbf{x}_1 \dots \mathbf{x}_n} = 1\}$ and the open affine subset

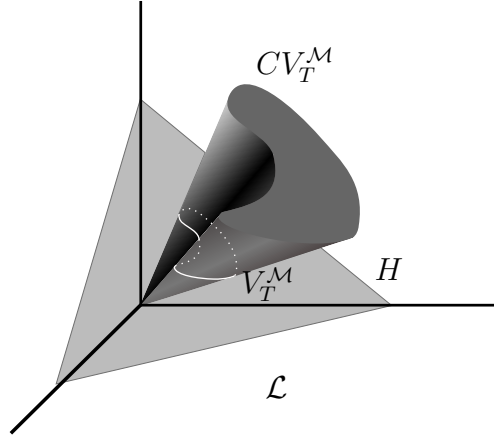


FIGURE 1.

$\{p = [p_{b_1 \dots b_1} : \dots : p_{b_k \dots b_k}] : \sum p_{x_1 \dots x_n} \neq 0\}$ of $\mathbb{P}^{k^n-1} = \mathbb{P}(\mathcal{L})$ (we use projective coordinates $[p_{b_1 \dots b_1} : \dots : p_{b_k \dots b_k}]$ to distinguish them from affine coordinates). The *projective phylogenetic variety* $\mathbb{P}V_T^{\mathcal{M}}$ associated to a phylogenetic tree T on W is the closure in $\mathbb{P}^{k^n-1} = \mathbb{P}(\mathcal{L})$ of the image of the stochastic parameterization $\phi_T^{\mathcal{M}}$ defined above.

The aim now is to study the relation between the above varieties. As it is usually easier to deal with a homogeneous parameterization and homogeneous polynomials, it will be useful to prove that $CV_T^{\mathcal{M}}$ is the cone over $\mathbb{P}V_T^{\mathcal{M}}$. This is known for some particular models (for instance, see [AR08] for a proof on the general Markov model) but as our definition of algebraic evolutionary model is quite general, we would like to prove it in its maximum generality.

We will use the following notation:

Notation 2.18. Given a matrix A , we will write $D_A = \text{diag}(A\mathbf{1})$. Given $p = (p_{x_1, \dots, x_n})_{x_1, \dots, x_n} \in \mathcal{L}$, write $\lambda(p) = \sum_{x_i \in B} p_{x_1, \dots, x_n}$. For example, we can write the hyperplane H above as $H = \{p \in \mathcal{L} : \lambda(p) = 1\}$.

The following proposition is an adaptation of [AR08, Proposition 1] to our models.

Proposition 2.19. *Let $\mathcal{M} = (W_0, \text{Mod})$ be a root-independent evolutionary model and let T be a trivalent n -leaf rooted tree on W evolving under \mathcal{M} . Fix an edge path γ in T with initial vertex the root of T and such that it passes through all vertices in the tree. Then there is a non-empty open set $U_\gamma \subset \text{Par}_{\mathcal{M}}(T)$ such that, if $p = \Psi_T^{\mathcal{M}}(\pi, \mathbf{A})$ with $(\pi, \mathbf{A}) \in U_\gamma$,*

- (i) there is a presentation $(\tilde{\pi}, \tilde{\mathbf{A}}) \in \text{Par}_{\mathcal{M}}(T)$ with \tilde{A}^e stochastic for every edge e in the tree and $p = \Psi_T^{\mathcal{M}}(\tilde{\pi}, \tilde{\mathbf{A}})$;
 (ii) there exists $q \in \text{Im } \phi_T^{\mathcal{M}}$ such that $p = \lambda(p)q$.

Proof. Given $A \in \text{Mod}$ and $\pi \in W_0$, write $\tilde{\pi} = A^t\pi$. Since the model is root-independent, the matrix $\tilde{A} = D_{\tilde{\pi}}^{-1}A^tD_{\pi}$ is still in Mod and all entries in $\tilde{\pi}$ are different from 0. In this case notice that \tilde{A} is stochastic:

$$\tilde{A}\mathbf{1} = (D_{\tilde{\pi}}^{-1}A^tD_{\pi})\mathbf{1} = D_{A^t\pi}^{-1}(A^t\pi) = \mathbf{1}.$$

The idea for the proof is to move the root from one vertex of the tree to another according to the edge path γ , replacing at each step some matrix A^e by a new matrix \tilde{A}^e which is stochastic. In order to construct this new matrix, some conditions must be required to the original presentation (π, \mathbf{A}) .

Write $\gamma = (e_1, e_2, \dots, e_m)$ for the ordered collection of edges in γ so that e_1 contains the root and the terminal vertex of e_{i-1} equals the initial vertex of e_i . Notice that the edge path may go twice through the same edge and, in this case, this edge will appear in the collection above with opposite orientation.

The root of T and each edge e_i provide a number of conditions on $\text{Par}_{\mathcal{M}}(T)$ as claimed, namely:

Root r : π has all its entries different from 0.

Edge e_1 : $\pi_1 := (A^{e_1})^t\pi$ has all its entries different from 0.

Then, write $A_1^{e_1} = D_{\pi_1}^{-1}(A^{e_1})^tD_{\pi}$.

Edge e_i , $2 \leq i \leq m$: $\pi_i := (A_{i-1}^{e_i})^t\pi_{i-1}$ has all its entries different from 0.

Then, write $A_{i+1}^{e_i} = D_{\pi_i}^{-1}(A_i^{e_i})^tD_{\pi_{i-1}}$.

Define U_{γ} as the set in $\text{Par}_{\mathcal{M}}(T)$ defined by all these conditions. It is a nonempty Zariski open set of $\text{Par}_{\mathcal{M}}(T)$. Moreover, for any $(\pi, \mathbf{A}) \in U_{\gamma}$, the inverse of D_{π} and the consecutive inverses of D_{π_i} are guaranteed to exist. Therefore, we apply Lemma 2.11 to move the root through the path γ and obtain a new presentation $(\tilde{\pi}, \tilde{\mathbf{A}})$ in \mathcal{M} satisfying the conditions of (i).

We can assume therefore that $p = \Psi_T^{\mathcal{M}}(\pi, \mathbf{A})$ with A^e stochastic for every edge in the tree. Since $\pi \in W_0$, we know that $\lambda := \mathbf{1}^t\pi \neq 0$. Define a new distribution for the root by $\tilde{\pi} = \frac{\pi}{\lambda} \in W_0$ and write $\tilde{\mathbf{A}} = \mathbf{A}$. We have that $(\tilde{\pi}, \tilde{\mathbf{A}}) \in s\mathcal{M}$ and we define $q = \phi_T^{\mathcal{M}}(\tilde{\pi}, \tilde{\mathbf{A}})$. We have to show that $p = \lambda q$ and $\lambda = \lambda(p)$.

Write q_{x_1, \dots, x_n} for its coordinates in the basis \mathcal{B} . Similarly, write p_{x_1, \dots, x_n} for the coordinates of p . Identifying the vectors $\mathbf{x} = \sum_{i=1}^k c_i b_i \in W$ with its

coordinates in the basis B , (c_1, c_2, \dots, c_k) , we write $\mathbf{x}^t M$ to mean $(c_1, c_2, \dots, c_k)M$. Applying (1), we obtain that have that

$$\begin{aligned} p_{\mathbf{x}_1 \dots \mathbf{x}_n} &= \sum_{\mathbf{x}_v \in B, v \in \text{Int}(T)} \pi_{\mathbf{x}_r} \prod_{i=1}^{2n-3} \mathbf{x}_{e_i^0}^t \tilde{A}^i \mathbf{x}_{e_i^1} \\ &= \lambda \sum_{\mathbf{x}_v \in B, v \in \text{Int}(T)} \tilde{\pi}_{\mathbf{x}_r} \prod_{i=1}^{2n-3} \mathbf{x}_{e_i^0}^t \tilde{A}^i \mathbf{x}_{e_i^1} = \lambda q_{\mathbf{x}_1 \dots \mathbf{x}_n}, \end{aligned}$$

the second equality by the definition of $\tilde{\pi}$. Moreover, since $\sum q_{\mathbf{x}_1, \dots, \mathbf{x}_n} = 1$, we infer that $\lambda = \sum_{\mathbf{x}_1 \dots \mathbf{x}_n} p_{\mathbf{x}_1 \dots \mathbf{x}_n}$, that is, $\lambda = \lambda(p)$. \square

Given a set $Z \subset \mathcal{L}$, denote by $\mathcal{I}(Z)$ the ideal of polynomials in $\mathbb{C}[\mathcal{L}] := \mathbb{C}[p_{\mathbf{x}_1, \dots, \mathbf{x}_n}]$ that vanish over Z . From the proof above, we state the following facts for future reference. We prove the following corollary relating the different phylogenetic varieties defined above.

Corollary 2.20. *Let $\mathcal{M} = (W_0, \text{Mod})$ be a root-independent evolutionary model and let T be a trivalent n -leaf tree on W evolving under \mathcal{M} . Then,*

- (a) $CV_T^{\mathcal{M}}$ equals the affine cone over the projective phylogenetic variety $\mathbb{P}V_T^{\mathcal{M}}$;
- (b) $\mathcal{I}(\text{Im } \Psi_T^{\mathcal{M}}) + (h) = \mathcal{I}(\text{Im } \phi_T^{\mathcal{M}})$, where $h = \sum p_{\mathbf{x}_1, \dots, \mathbf{x}_n} - 1$;
- (c) $V_T^{\mathcal{M}} = CV_T^{\mathcal{M}} \cap H$.

Consequence (a) was proved by Allman and Rhodes for the general Markov model (see [AR08, Proposition 1]).

Proof. (a) Since $\text{Im } \Psi_T^{\mathcal{M}}$ is a cone, the ideal of polynomials vanishing on it has to be homogenous. Now, by virtue of Proposition 2.19 we know that a homogenous polynomial vanishes on $\text{Im } \Psi_T^{\mathcal{M}}$ if and only if it vanishes on $\text{Im } \Phi_T^{\mathcal{M}}$. It follows immediately that $CV_T^{\mathcal{M}}$, which is the variety defined by all these polynomials in \mathcal{L} , equals the affine cone over $\mathbb{P}V_T^{\mathcal{M}}$.

(b) One inclusion is easy. To prove the other inclusion, let $F \in \mathcal{I}(\text{Im } \Phi_T^{\mathcal{M}})$, so that F vanishes on $V_T^{\mathcal{M}}$. If d is the degree of F , write $F = F_m + F_{m+1} + \dots + F_d$, where every F_j is a homogenous polynomial of degree $j \leq d$. Let $p \in \text{Im}(\Psi_T^{\mathcal{M}})$ and, using Proposition 2.19 write $p = \lambda(p)q$, where $q \in \text{Im}(\phi_T^{\mathcal{M}})$. Then, we have

$$\begin{aligned} 0 = F(q) &= F_m \left(\frac{p}{\lambda(p)} \right) + F_{m+1} \left(\frac{p}{\lambda(p)} \right) + \dots + F_d \left(\frac{p}{\lambda(p)} \right) = \\ &= \frac{F_m(p)}{\lambda(p)^m} + \frac{F_{m+1}(p)}{\lambda(p)^{m+1}} + \dots + \frac{F_d(p)}{\lambda(p)^d}. \end{aligned}$$

On the other hand, the equation of H is $h = 0$ where $h(p) = \lambda(p) - 1$. That is, $\lambda(p) = h(p) + 1$. Replacing this in the above equation and multiplying by $\lambda(p)^d$, we obtain that the polynomial

$$\bar{F}(p) = F_m(p)(h(p) + 1)^{d-m} + F_{m+1}(p)(h(p) + 1)^{d-m-1} \dots + F_d(p) = 0$$

is identically zero on $\text{Im}\Psi_T^{\mathcal{M}}$, that is, $\bar{F} \in \mathcal{I}(\text{Im}\Psi_T^{\mathcal{M}})$. This polynomial has the form $\bar{F} = F + hQ$ for some polynomial $Q \in \mathbb{C}[\mathcal{L}]$. From this, we immediately have $F = \bar{F} - hQ \in \mathcal{I}(\text{Im}\Psi_T^{\mathcal{M}}) + (h)$ and we are done.

(c) follows directly by taking the affine varieties defined by the ideals in (b). \square

The previous Corollary implies that $\dim CV_T^{\mathcal{M}} = \dim \mathbb{P}V_T^{\mathcal{M}} + 1$, and if $p = (p_{A\dots A}, \dots, p_{T\dots T})$ belongs to $CV_T^{\mathcal{M}}$, then $q := [p_{A\dots A} : \dots : p_{T\dots T}]$ belongs to $\mathbb{P}V_T^{\mathcal{M}}$. Moreover, if $\lambda := \sum p_{x_1\dots x_n} \neq 0$, then $q = [\frac{p_{A\dots A}}{\lambda} : \dots : \frac{p_{T\dots T}}{\lambda}]$ and $(\frac{p_{A\dots A}}{\lambda}, \dots, \frac{p_{T\dots T}}{\lambda})$ is a point in the affine stochastic phylogenetic variety $V_T^{\mathcal{M}}$.

3. THE SPACE OF PHYLOGENETIC MIXTURES

In phylogenetics, the hypothesis that the sites of an alignment are independent and identically distributed is often used in the most simple models. When one removes the assumption “identically distributed” and replaces it by “distributed according to the same evolutionary model” then one obtains a phylogenetic mixture. Here we introduce a phylogenetic mixture from the algebraic point of view.

Definition 3.1. 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 $p^i \in \text{Im}(\Psi_{T_i}^{\mathcal{M}})$, $T_i \in \mathbb{T}_n$ and $\alpha_i \in \mathbb{C}$. As $\Psi_{T_i}^{\mathcal{M}}$ is a homogeneous map, phylogenetic mixtures are actually vectors of the form $\sum_{i=1}^m \check{p}^i$, where $\check{p}^i \in \text{Im}(\Psi_{T_i}^{\mathcal{M}})$.

Note that on a phylogenetic mixture we allow some (or all) tree topologies T_i to be the same. Therefore, the widely used discrete Gamma-rates or any type of rate variability across sites are instances of phylogenetic mixtures (we refer to the book [SS03] for an introduction to these concepts.)

We denote by $\mathcal{D}_{\mathcal{M}} \subset \mathcal{L}$ the set of all phylogenetic mixtures (on any number of classes) under the algebraic evolutionary model \mathcal{M} and by $\mathcal{D}_{\mathcal{M}}^m$ the set of all phylogenetic mixtures on m -classes.

When we restrict to matrices whose rows sum to one so that we consider the parameterization $\phi_T^{\mathcal{M}}$, one has to restrict the phylogenetic mixtures to points of the form

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

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

The following result was proven by Matsen, Mossel and Steel in [MMS08] for the two state random cluster model.

Lemma 3.2. *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, the space $\mathcal{D}_{s\mathcal{M}}$ is a linear variety of the affine space \mathcal{L} contained in the hyperplane H .*

Proof. $\mathcal{D}_{\mathcal{M}}$ is a \mathbb{C} -vector space by definition.

In order to prove that $\mathcal{D}_{s\mathcal{M}}$ is a linear variety, let q_0 be any point in $\mathcal{D}_{s\mathcal{M}}$, so that $q_0 = \sum_{i=1}^m \alpha_i q^i$ with $q^i \in \text{Im}(\phi_{T_i}^{\mathcal{M}})$, $i = 1, \dots, m$, and $\sum_i \alpha_i = 1$. Then we can write

$$\mathcal{D}_{s\mathcal{M}} = q_0 + F, \quad \text{where} \quad F = \{\overrightarrow{q_0 q} \mid q \in \mathcal{D}_{s\mathcal{M}}\}.$$

We only have to show that F is a \mathbb{C} -vector space:

- 1) Let $v = \overrightarrow{q_0 \hat{q}}$ be a vector in F , then $\lambda v = \overrightarrow{q_0 q'}$ where $q' = q_0 + \lambda \overrightarrow{q_0 \hat{q}}$. This last point is in $\mathcal{D}_{s\mathcal{M}}$: if $q = \sum_{j=1}^l \beta_j \hat{q}^j$ with $\sum_j \beta_j = 1$, then $q' = (1 - \lambda) \sum_{i=1}^m \alpha_i q^i + \lambda \sum_{j=1}^l \beta_j \hat{q}^j$ and the scalar coefficients sum to one $(1 - \lambda) \sum_i \alpha_i + \lambda \sum_j \beta_j = (1 - \lambda) + \lambda = 1$. Therefore λv is in F .
- 2) Let $v_1 = \overrightarrow{q^0 \hat{q}^1}$ and $v_2 = \overrightarrow{q^0 \check{q}^2}$ be two vectors in F ,

$$q^1 = \sum_j \beta_j \hat{q}_j \quad \text{with} \quad \sum_j \beta_j = 1,$$

$$q^2 = \sum_k \gamma_k \check{q}_k \quad \text{with} \quad \sum_k \gamma_k = 1,$$

then $v_1 + v_2 = \overrightarrow{q^0 q'}$ with $q' = \sum_j \beta_j \hat{q}_j + \sum_k \gamma_k \check{q}_k - \sum_i \alpha_i q_i$, and all coefficients together sum to one: $\sum_j \beta_j + \sum_k \gamma_k - \sum_i \alpha_i = 1$.

□

Remark 3.3. By virtue of the previous lemma, $\mathcal{D}_{\mathcal{M}}$ is an algebraic variety that contains $\text{Im}\Psi_T^{\mathcal{M}}$ for any tree T and therefore, it also contains $CV_T^{\mathcal{M}}$. It follows

that $\mathcal{D}_{\mathcal{M}}$ equals the set of points of the form $p = \sum p^i$ where $p^i \in CV_{T_i}^{\mathcal{M}}$. Similarly, $\mathcal{D}_{s\mathcal{M}}$ equals the set of points of the form $q = \sum \alpha_i q_i$, where $q_i \in V_{T_i}^{\mathcal{M}}$ and $\sum_i \alpha_i = 1$.

For technical reasons needed in the next result, we introduce the following spaces:

Definition 3.4. Define $\overline{\mathcal{D}_{\mathcal{M}}^m}$ as the set of points p of the form $p = \sum_{i=1}^m p^i$ where $p^i \in CV_{T_i}^{\mathcal{M}}$, and $\overline{\mathcal{D}_{s\mathcal{M}}^m}$ as the set of points q of the form $q = \sum_{i=1}^m \alpha_i q^i$ where $q^i \in V_{T_i}^{\mathcal{M}}$ and $\sum_{i=1}^m \alpha_i = 1$.

Lemma 3.5. *The following equalities hold*

$$\begin{aligned} (a) \quad \overline{\mathcal{D}_{s\mathcal{M}}^m} &= \overline{\mathcal{D}_{\mathcal{M}}^m} \cap H \\ (b) \quad \mathcal{D}_{s\mathcal{M}} &= \mathcal{D}_{\mathcal{M}} \cap H. \end{aligned}$$

Proof. (a) Let $q \in \overline{\mathcal{D}_{s\mathcal{M}}^m}$. Then, we can write $q = \sum_{i=1}^m \alpha_i q^i$ for some $q^i \in V_{T_i}^{\mathcal{M}}$ and $\sum \alpha_i = 1$. Clearly, $q \in \overline{\mathcal{D}_{\mathcal{M}}^m}$. Moreover, $\lambda(q) = \sum_i \alpha_i \lambda(q^i) = \sum_i \alpha_i = 1$. 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 Proposition 2.19 to each p^i to get $p^i = \lambda(p^i)q_i$ for some $q_i \in V_{T_i}^{\mathcal{M}}$. Then, we have

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

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 $p \in \overline{\mathcal{D}_{s\mathcal{M}}^m}$.

(b) can be proven using (a) and Remark 3.3. □

4. THE SPACE OF PHYLOGENETIC MIXTURES FOR EQUIVARIANT EVOLUTIONARY MODELS

This section will be devoted to give a precise description of the space $\mathcal{D}_{\mathcal{M}}$ for the equivariant models \mathcal{M} listed in 2.5. Thus, we will assume that $B = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$, $k = 4$ and $W = \langle B \rangle_{\mathbb{C}}$. From now on, n is fixed and $\mathcal{L} = \otimes^n W$.

Let $G \leq \mathfrak{S}_4$ be a permutation group. We consider the restriction to G of the *defining* representation

$$(2) \quad \rho : \mathfrak{S}_4 \rightarrow GL(W)$$

given by the permutation of the elements of B . This representation induces a G -module structure on W by setting

$$g \cdot \mathbf{x} := \rho(g)(\mathbf{x}) \in W.$$

In fact, ρ induces a G -module structure on $\mathcal{L} = \otimes^n W$ by setting

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

and extends by linearity. According to Notation 2.9, if $\mathbf{X} \in B$ and $g \in G$, $g\mathbf{X}$ will stand for the action of g on \mathbf{X} as introduced above. From now on, the space \mathcal{L} will be implicitly considered as a G -module with this action.

Definition 4.1. Given a set of taxa $[n]$, a G -tensor on $[n]$ is an $[n]$ -tensor invariant by the action defined in (3). The set of G -tensors will be denoted by \mathcal{L}^G .

The following Theorem describes the set of phylogenetic mixtures for equivariant models in an easy way.

Theorem 4.2. *If \mathcal{M} is one of the equivariant evolutionary models JC69, K80, K81, SSM or GMM, then the space of phylogenetic mixtures $\mathcal{D}_{\mathcal{M}}$ coincides with $\mathcal{L}^{G_{\mathcal{M}}}$ and $\mathcal{D}_{s_{\mathcal{M}}} = \mathcal{L}^{G_{\mathcal{M}}} \cap H$.*

This theorem allows one to identify the set of all phylogenetic mixtures $\mathcal{D}_{\mathcal{M}}$ with $\mathcal{L}^{G_{\mathcal{M}}}$, which is a vector subspace of \mathcal{L} whose linear equations are easy to describe, as we will see afterwards in this section. In other words, $\mathcal{L}^{G_{\mathcal{M}}}$ is the space where data coming from any mixture of trees evolving under model \mathcal{M} lies. One can therefore use $\mathcal{L}^{G_{\mathcal{M}}}$ to select the most suitable model for given data. This has been studied in the paper [KDGC11] by the first and third author jointly with M. Drton and R. Guigó.

Proof of Theorem 4.2. In Lemma 3.2 we proved that $\mathcal{D}_{\mathcal{M}}$ is a vector subspace of \mathcal{L} . Moreover, as we are considering equivariant models, we have $\text{Im}(\Psi_T^{\mathcal{M}}) \subset \mathcal{L}^{G_{\mathcal{M}}}$ for any tree T (see Lemma 4.3 of [DK09]) and hence $\mathcal{D}_{\mathcal{M}}$ is contained in the vector subspace $\mathcal{L}^{G_{\mathcal{M}}}$.

In order to show that $\mathcal{L}^{G_{\mathcal{M}}} = \mathcal{D}_{\mathcal{M}}$ it remains to prove that there does not exist any hyperplane Π containing $\mathcal{D}_{\mathcal{M}}$ and not containing $\mathcal{L}^{G_{\mathcal{M}}}$. If such a hyperplane existed, then it would contain, in particular, all points in $\text{Im} \Psi_T^{\mathcal{M}}$ for any tree topology T . As Π is an algebraic variety, this implies that Π contains $CV_T^{\mathcal{M}}$ for any tree topology T .

It is enough to prove that, for the equivariant models considered here, there are no homogeneous linear polynomials vanishing on all tree topologies, except the linear equations vanishing on $\mathcal{L}^{G_{\mathcal{M}}}$. This is already known in the literature:

for G corresponding to **GMM** this was proven in Allman-Rhodes [AR04]; for the Strand symmetric model this is in [CS05]; for **JC69,K80,K81** this appears in [SS05] (for **JC69** and **K80** there are other linear relations but they correspond to phylogenetic invariants, i.e. they are equations that vanish on $\Psi_T^{\mathcal{M}}$ for a particular tree topology T but not for all topologies). The main result in [DK09] comprises all these results.

The equality $\mathcal{D}_{s,\mathcal{M}} = \mathcal{L}^{G,\mathcal{M}} \cap H$ follows immediately from Lemma 3.5 and the first assertion in this theorem. \square

4.1. Equations for the space $\mathcal{L}^{G,\mathcal{M}}$. Our purpose now is to compute the dimension of $\mathcal{L}^{G,\mathcal{M}}$ where \mathcal{M} is one of the equivariant models listed in Definition 2.5, as well as to obtain a set of independent linear equations defining this space. To this aim we need to recall some definitions and facts in group theory and group representation theory.

Let $G \leq \mathfrak{S}_4$ be a permutation group. Given an element $g \in G$, the *conjugacy class* of g is $C(g) = \{h^{-1}gh : h \in G\}$. If $g_1, g_2 \in G$, then it is easy to see that either $C(g_1) = C(g_2)$ or $C(g_1) \cap C(g_2) = \emptyset$. 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 $\sum_{i=1}^s |C_i| = |G|$. Write χ^n for the character of G associated to the defining representation $G \rightarrow GL(\otimes^n W)$. Recall that $\chi^n(g_1) = \chi^n(g_2)$ whenever g_1 and g_2 lie in the same conjugacy class, so that we represent χ^n by a s -tuple (t_1, \dots, t_s) where $t_i = \chi^n(g)$ for any $g \in C_i$.

Example 4.3. (cf. [CFS11]) For the equivariant models listed in Definition 2.5, we have the following table (denote by e the trivial permutation of \mathfrak{S}_4):

$G \leq \mathfrak{S}_4$	\mathcal{M}	representants of conj. classes	$\mathcal{C}(G)$	(t_1, \dots, t_s)
$\langle\langle \text{(AT)(CG)} \rangle\rangle$	SSM	$\{e, \text{(AT)(CG)}\}$	(1, 1)	$(4^n, 0)$
$\langle\langle \text{(AC)(GT)}, \text{(AG)(CT)} \rangle\rangle$	K81	$\{e, \text{(AT)(CG)}, \text{(AC)(GT)}, \text{(AG)(CT)}\}$	(1, 1, 1, 1)	$(4^n, 0, 0, 0)$
$\langle\langle \text{(ACGT)}, \text{(AG)} \rangle\rangle$	K80	$\{e, \text{(AC)(GT)}, \text{(AG)(CT)}, \text{(ACGT)}, \text{(AG)}\}$	(1, 2, 1, 2, 2)	$(4^n, 0, 0, 0, 2^n)$
\mathfrak{S}_4	JC69	$\{e, \text{(AC)(GT)}, \text{(ACGT)}, \text{(AG)}, \text{(ACG)}\}$	(1, 3, 6, 6, 8)	$(4^n, 0, 0, 2^n, 1)$

Let $\Omega_G = \{\omega\}_{i=1, \dots, t}$ be a set of the irreducible characters of G , where ω_1 stands for the trivial character. Marshche's Theorem applied to the action of G described in (3) states that there is a decomposition of $\otimes^n W$ into its isotypic components:

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

where each $(\otimes^n W)[\omega_i]$ is isomorphic to a number of copies of the irreducible representation N_i associated to ω_i , $(\otimes^n W)[\omega_i] \cong N_i \otimes \mathbb{C}^{m_i(n)}$ for some positive

integer $m_i(n)$, called the *multiplicity of $\otimes^n W$ relative to ω_i* . Moreover, the set Ω_G forms an orthonormal basis of the space of characters relative to the inner product defined by

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

Proposition 4.4. *We have*

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

Proof. Let \mathcal{M} be either SSM, K81, K80 or JC69. First of all, notice that the space of $G_{\mathcal{M}}$ -tensors is just the isotypic component of $\otimes^n W$ associated to the trivial representation, or equivalently, to the trivial character ω_1 :

$$\mathcal{L}^{\mathcal{M}} = (\otimes^n W)[\omega_1].$$

Since the dimension of the trivial representation is one, it follows that the dimension of $\mathcal{L}^{\mathcal{M}}$ is precisely the multiplicity $m_1(n)$, that is, the number of times the trivial representation appears in the decomposition of $\otimes^n W$ into isotypic components. This multiplicity equals

$$m_1(n) = \langle \chi^n, \omega_1 \rangle = \frac{1}{|G|} \sum_{g \in G} \chi^n(g) \omega_1(g) = \frac{1}{|G|} \sum_{i=0}^s |C_i| t_i,$$

where in the last equality we group the elements of G according to their conjugacy class. We apply this formula to the different groups described in Example 4.3 and the result follows. \square

Our next goal is to provide a set of independent linear equations for $\mathcal{L}^{G_{\mathcal{M}}}$. From now on, we will use the notation introduced in 2.9 and we add the following notation before stating the main result.

Notation 4.5. We will consider the following subsets of $\mathcal{B} = B^n$:

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

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

We will adopt multiplicative notation for the n -words in the alphabet B . For instance, we will write \mathbf{C}^l to mean the word and $\underbrace{\mathbf{C} \dots \mathbf{C}}_l$ and $(\mathbf{A}^l)(\mathbf{G}^m)\mathbf{x}_{l+m+1} \dots \mathbf{x}_n$ to mean $\underbrace{\mathbf{A} \dots \mathbf{A}}_l \underbrace{\mathbf{G} \dots \mathbf{G}}_m \mathbf{x}_{l+m+1} \dots \mathbf{x}_n$, where $\mathbf{x}_{l+m+1}, \dots, \mathbf{x}_n$ represent any possible choice of letters.

The main result of this section is the following:

Theorem 4.6. *A set of linearly independent equations $\mathbb{E}^{\mathcal{M}}$ for $\mathcal{L}^{\mathcal{G}\mathcal{M}}$ is given by*

$$\begin{aligned} \mathbb{E}^{\text{SSM}} &: p_{\mathbf{X}} = p_{(\text{AT})(\text{CG})\mathbf{X}} \text{ where } \mathbf{X} \text{ has } \mathbf{x}_1 \in \{\mathbf{A}, \mathbf{C}\}; \\ \mathbb{E}^{\text{K81}} &: \text{the equations in } \mathbb{E}^{\text{SSM}}, \text{ together with} \end{aligned}$$

$$p_{\mathbf{X}} = p_{(\text{AC})(\text{GT})\mathbf{X}},$$

where \mathbf{X} has $\mathbf{x}_1 = \mathbf{A}$;

$$\mathbb{E}^{\text{K80}} : \text{the equations in } \mathbb{E}^{\text{K81}}, \text{ together with}$$

$$p_{\mathbf{X}} = p_{(\text{AG})\mathbf{X}},$$

where $\mathbf{X} \in \mathcal{B} \setminus \mathcal{B}_{\text{AC|GT}}$ has $\mathbf{x}_1 = \mathbf{A}$, and if \mathbf{T} appears in \mathbf{X} , there is some \mathbf{C} in a preceding position;

$$\mathbb{E}^{\text{JC69}} : \text{the equations in } \mathbb{E}^{\text{K80}}, \text{ together with}$$

$$p_{\mathbf{X}} = p_{(\text{AT})\mathbf{X}},$$

where $\mathbf{X} \in \mathcal{B}_{\text{AC|GT}} \setminus \mathcal{B}_0$ has the form $(\mathbf{A}^l)(\mathbf{C}^m)\mathbf{x}_{l+m+1} \dots \mathbf{x}_n$; and equations

$$p_{\mathbf{X}} = p_{(\text{AC})\mathbf{X}} \quad \text{and} \quad p_{\mathbf{X}} = p_{(\text{AT})\mathbf{X}}$$

where $\mathbf{X} \in \mathcal{B} \setminus \mathcal{B}_2$ has the form $(\mathbf{A}^l)(\mathbf{C}^m)\mathbf{x}_{l+m+1} \dots \mathbf{x}_n$ and, if \mathbf{T} appears in \mathbf{X} , there is some \mathbf{G} in a preceding position.

The number of equations added in each case is:

$$\begin{aligned} \text{SSM} &: 2^{2n-1}; \\ \text{K81} &: 2^{2n-2}; \\ \text{K80} &: 2^{2n-3} - 2^{n-2}; \text{ and} \\ \text{JC69} &: 2^{n-1} - 1 + 2\left(\frac{2^{2n-3}+1}{3} - 2^{n-2}\right). \end{aligned}$$

In order to prove this theorem we need a few technical results.

Lemma 4.7. *If $G = \langle g_1, \dots, g_t \rangle$, then $\mathcal{L}^G = \bigcap_{i=1}^t \mathcal{L}^{\langle g_i \rangle}$.*

Proof. One inclusion is straightforward. We prove the other. Let $p \in \bigcap_{i=1}^s \mathcal{L}^{\langle g_i \rangle}$, so we have $g_i p = p$ for any i (and in particular, $g_i^{-1} p = p$). Any element of G can be written as $g = g_{i_1}^{m_1} \dots g_{i_r}^{m_r}$ with $m_i \neq 0$. The invariance of p under all the g_i and g_i^{-1} completes the proof. \square

As a consequence of this lemma, we obtain that a system of linear equations for \mathcal{L}^G is obtained from a system of generators of G : given a point $p \in \mathcal{L}$, we have that

$$p \in \mathcal{L}^G \quad \Leftrightarrow \quad p_{g\mathbf{x}} = p_{\mathbf{x}}, \quad \forall g \in G, \forall \mathbf{x} \in \mathcal{B}.$$

If H is a subgroup of G , we take $H \setminus G = \{Hg : g \in G\}$ for the set of right cosets of H in G , $Hg = \{hg : h \in H\}$. By Lagrange's theorem, we know that $|H \setminus G| = |G|/|H|$. Moreover, if $[G : H]$ is the *index* of H in G and $\{g_1, \dots, g_{[G:H]}\}$ is a transversal of $H \setminus G$, we have a partition of G

$$(6) \quad G = \bigcup_{i=1}^{[G:H]} Hg_i.$$

The set $H \setminus G$ can be understood as a single G -orbit with the natural action of G on it.

Example 4.8. For the models of Example 4.3, we have

- (i) $[G_{\text{SSM}} : \langle e \rangle] = 2$; a transversal of $\langle e \rangle \setminus G_{\text{SSM}}$ is $\{e, (\text{AT})(\text{CG})\}$.
- (ii) $[G_{\text{K81}} : G_{\text{SSM}}] = 2$; a transversal of $G_{\text{SSM}} \setminus G_{\text{K81}}$ is $\{e, (\text{AC})(\text{GT})\}$.
- (iii) $[G_{\text{K80}} : G_{\text{K81}}] = 2$; a transversal of $G_{\text{K81}} \setminus G_{\text{K80}}$ is $\{e, (\text{AG})\}$.
- (iv) $[G_{\text{JC69}} : G_{\text{K80}}] = 3$; a transversal of $G_{\text{K80}} \setminus G_{\text{JC69}}$ is $\{e, (\text{AC}), (\text{AT})\}$.

Notation. We write $\{\mathbf{X}\}_G$ (or even $\{\mathbf{X}\}_{\mathcal{M}_G}$) for the orbit of $\mathbf{X} \in \mathcal{B}$ under the action of G : $\{\mathbf{X}\}_G = \{g\mathbf{X} : g \in G\}$.

Lemma 4.9. *Let g_1, \dots, g_m be a transversal of $H \setminus G$. For every $\mathbf{X} \in \mathcal{B}$, we have*

$$\{\mathbf{X}\}_G = \bigcup_{i=1, \dots, m} \{g_i \mathbf{X}\}_H.$$

Proof. Apply the decomposition (6) to element \mathbf{X} . \square

Lemma 4.10. *Let $\mathbf{X} \in \mathcal{B}$. Then,*

SSM: $\{\mathbf{X}\}_{\text{SSM}} = \{\mathbf{X}, (\text{AT})(\text{CG})\mathbf{X}\}$ and there are 2^{2n-1} different orbits.

- K81: $\{X\}_{K81} = \{X\}_{SSM} \cup \{(AC)(GT)X\}_{SSM}$ has cardinality 4 and there are 2^{2n-2} different orbits.
- K80: \circ If $X \in \mathcal{B}_{AG|CT}$ then $\{X\}_{K80} = \{X\}_{K81}$ has cardinality 4 and there are 2^{n-1} different orbits;
- \circ if $X \in \mathcal{B} \setminus \mathcal{B}_{AG|CT}$, then $\{X\}_{K80} = \{X\}_{K81} \cup \{(AG)X\}_{K81}$ has cardinality 8 and there are $2^{2n-3} - 2^{n-2}$ different orbits.
- JC69: \circ If $X \in \mathcal{B}_0$ then $\{X\}_{JC69} = \{X\}_{K80}$ has cardinality 4 and there is only one orbit;
- \circ if $X \in \mathcal{B}_{AC|GT} \setminus \mathcal{B}_0$ then $\{X\}_{JC69} = \{X\}_{K80} \cup \{(AT)X\}_{K80}$ has cardinality 12 and there are $2^{n-1} - 1$ different orbits; moreover, the union of such orbits cover the whole $\mathcal{B}_2 \setminus \mathcal{B}_0$.
- \circ if $X \in \mathcal{B} \setminus \mathcal{B}_2$ then $\{X\}_{JC69} = \{X\}_{K80} \cup \{(AC)X\}_{K80} \cup \{(AT)X\}_{K80}$ has cardinality 24 and there are $\frac{1}{3}(2^{2n-3} + 1) - 2^{n-2}$ different orbits.

We can summarize this result in the following table:

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

where \dots means the set on the left and " means the set on the top.

Proof. The idea of the proof is to systematically apply Lemma 4.9 to describe the orbits of the elements $X \in \mathcal{B}$ under the action of the groups considered. SSM and K81 are straightforward and are left to the reader.

K80: Applying Lemma 4.9, we obtain that

$$\{X\}_{K80} = \{X\}_{K81} \cup \{(AG)X\}_{K81}.$$

If $X \in \mathcal{B}_{AG|CT}$, then $\{(AG)X\}_{K81} = \{X\}_{K81}$ and $\{X\}_{K80}$ has cardinality 4. The number of such orbits is

$$\frac{|\mathcal{B}_{AG|CT}|}{4} = 2^{n-1}.$$

If $X \notin \mathcal{B}_{AG|CT}$, then $\{(AG)X\}_{K81} \neq \{X\}_{K81}$, so $\{X\}_{K80}$ has cardinality 8. The number of such orbits is

$$\frac{|\mathcal{B} \setminus \mathcal{B}_{AG|CT}|}{8} = 2^{2n-3} - 2^{n-2}.$$

JC69: Lemma 4.9 applies to give

$$\{X\}_{JC69} = \{X\}_{K80} \cup \{(AC)X\}_{K80} \cup \{(AT)X\}_{K80}.$$

- (a) If $\mathbf{X} \in \mathcal{B}_0$, then $\{(\mathbf{AC})\mathbf{X}\}_{\mathbf{K80}} = \{(\mathbf{AT})\mathbf{X}\}_{\mathbf{K80}} = \{\mathbf{X}\}_{\mathbf{K80}}$, so $\{\mathbf{X}\}_{\mathbf{JC69}}$ has 4 elements. The number of such orbits is

$$|\mathcal{B}_0|/4 = 1.$$

- (b) If $\mathbf{X} \in \mathcal{B}_{\mathbf{AC|GT}} \setminus \mathcal{B}_0$, then $(\mathbf{AT})\mathbf{X} \in \mathcal{B}_{\mathbf{AG|CT}}$ and $\{(\mathbf{AC})\mathbf{X}\}_{\mathbf{K80}} = \{\mathbf{X}\}_{\mathbf{K80}}$ has cardinality 8. Therefore, $\{\mathbf{X}\}_{\mathbf{JC69}} = \{(\mathbf{AT})\mathbf{X}\}_{\mathbf{K80}} \cup \{\mathbf{X}\}_{\mathbf{K80}}$ has cardinality $4 + 8 = 12$. The number of such orbits is

$$|\mathcal{B}_{\mathbf{AC|GT}} \setminus \mathcal{B}_0|/4 = 2^{n-1} - 1.$$

Moreover, the number of words involved in such orbits is

$$12(2^{n-1} - 1) = 3 \cdot 2^{n+1} - 12,$$

which is the cardinality of $\mathcal{B}_2 \setminus \mathcal{B}_0$.

- (c) Finally, if $\mathbf{X} \notin \mathcal{B}_2$, then the three orbits $\{(\mathbf{AC})\mathbf{X}\}_{\mathbf{K80}}$, $\{(\mathbf{AT})\mathbf{X}\}_{\mathbf{K80}}$ and $\{\mathbf{X}\}_{\mathbf{K80}}$ have 8 elements each and are disjoint. Thus, we obtain that

$$\{\mathbf{X}\}_{\mathbf{JC69}} = \{\mathbf{X}\}_{\mathbf{K80}} \cup \{(\mathbf{AC})\mathbf{X}\}_{\mathbf{K80}} \cup \{(\mathbf{AT})\mathbf{X}\}_{\mathbf{K80}}$$

has 24 elements. The number of such orbits is

$$\frac{|\mathcal{B} \setminus \mathcal{B}_2|}{24} = \frac{4^n - 3 \cdot 2^{n+1} + 8}{24} = \frac{2^{2n-3} + 1}{3} - 2^{n-2}.$$

This proves the claim. □

Remark 4.11. Notice that given a subgroup G of \mathfrak{S}_4 , every orbit $o = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ described above provides a G -tensor (a tensor invariant under the action of G) defined by

$$\Sigma(o) = \sum_{i=1}^m \mathbf{x}_i.$$

All these tensors are linearly independent, since each orbit involves different vectors of \mathcal{B} . It follows that all together they provide a basis for \mathcal{L}^G .

Now, we proceed to prove Theorem 4.6.

Proof of Theorem 4.6. In all these cases, the equations are obtained by taking the corresponding transversals given by Example 4.8. Assume we have computed a system of equations for the equivariant model associated with some subgroup $H \leq G$. By virtue of the previous lemmas we only need to care about the permutations added to H to generate G . Lemma 4.9 says that the new G -orbits result on the glueing of some H -orbits by the action of the new permutations

added. Therefore, the new equations to add are obtained by taking a transversal $\{g_1 = e, \dots, g_{[G:H]}\}$ of $H \setminus G$:

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

To avoid repetitions of equations, we have to choose a single element for every G -orbit. Notice that it may happen that for some $\mathbf{X} \in \mathcal{B}$, $\{g_i \mathbf{X}\}_H = \{g_j \mathbf{X}\}_H$ for $i \neq j$. In that 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 restriction. We have to take into account this possibility in order to obtain a minimal set of equations. That they form a minimal system of equations will follow from their cardinality and the dimension computation of Proposition 4.4.

SSM: As G_{SSM} is generated by (AT)(CG), a set of equations defining \mathcal{L}^{SSM} is

$$\{p_{\mathbf{X}} = p_{(\text{AT})(\text{CG})\mathbf{X}} : \mathbf{X} \in \mathcal{B}\}.$$

Each SSM-orbit provides a single equation. In order to avoid repetitions of equations, we take \mathbf{X} with $\mathbf{x}_1 \in \{\mathbf{A}, \mathbf{C}\}$. All together, we obtain 2^{2n-1} equations.

K81: As $\{e, (\text{AC})(\text{GT})\}$ is a transversal of $G_{\text{K81}} \setminus G_{\text{SSM}}$,

$$\{p_{\mathbf{X}} = p_{(\text{AC})(\text{GT})\mathbf{X}} : \mathbf{X} \in \mathcal{B}\}.$$

As above, each K81-orbit gives rise to a single equation. To avoid repetitions, we restrict to \mathbf{X} with $\mathbf{X}_1 = \mathbf{A}$. Therefore, we are adding 2^{2n-2} equations.

K80: As $\{e, (\text{AG})\}$ is a transversal of $G_{\text{K80}} \setminus G_{\text{K81}}$, we add the equations

$$\{p_{\mathbf{X}} = p_{(\text{AG})\mathbf{X}} : \mathbf{X} \in \mathcal{B}\}.$$

In order to decide whether we are actually adding new equations, we use Lemma 4.10. If $\mathbf{X} \in \mathcal{B}_{\text{AG|CT}}$, we know that $\{\mathbf{X}\}_{\text{K80}} = \{\mathbf{X}\}_{\text{K81}}$ and thus these orbits do not give rise to new equations. On the other hand, every orbit $\{\mathbf{X}\}_{\text{K80}}$ where $\mathbf{X} \notin \mathcal{B}_{\text{AG|CT}}$, provides a single equation. To avoid repetitions, we take \mathbf{X} with $\mathbf{X}_1 = \mathbf{A}$ and if T appears in \mathbf{X} , there is some C in a preceding position. Since $\mathbf{X} \notin \mathcal{B}_{\text{AG|CT}}$, the existence and unicity of such an element in every G_{K80} -orbit is guaranteed. We are adding

$$(1 - 1) \times 2^{n-1} + (2 - 1) \times (2^{2n-3} - 2^{n-2}) = 2^{2n-3} - 2^{n-2}$$

new equations.

JC69: As $\{e, (\text{AC}), (\text{AT})\}$ is a transversal of $G_{\text{JC69}} \setminus G_{\text{K80}}$, we add the equations

$$\{p_{\mathbf{X}} = p_{(\text{AC})\mathbf{X}} : \mathbf{X} \in \mathcal{B}\} \cup \{p_{\mathbf{X}} = p_{(\text{AT})\mathbf{X}} : \mathbf{X} \in \mathcal{B}\}.$$

In what follows we use Lemma 4.10 to get rid of redundant equations. If $\mathbf{X} \in \mathcal{B}_0$, then $\{\mathbf{X}\}_{\text{K80}} = \{(\text{AC})\mathbf{X}\}_{\text{K80}} = \{(\text{AT})\mathbf{X}\}_{\text{K80}}$, so we obtain nothing new in this case.

If $\mathbf{X} \in \mathcal{B}_{\text{AG|CT}} \setminus \mathcal{B}_0$, we add the equations

$$p_{\mathbf{X}} = p_{(\text{AT})\mathbf{X}}.$$

To avoid repetitions, we take \mathbf{X} of the form $(\mathbf{A}^l)(\mathbf{C}^m)\mathbf{x}_{l+m+1} \dots \mathbf{x}_n$, where $l, m \geq 1$: we are adding $2^{n-1} - 1$ new equations.

By Lemma 4.10, if $\mathbf{X} \in \mathcal{B}_{\text{AG|CT}} \cup \mathcal{B}_{\text{AT|CG}} \setminus \mathcal{B}_0$, then the corresponding JC69-orbit contains elements of $\mathcal{B}_{\text{AG|CT}}$ and therefore these orbits do not provide new equations.

Finally, if $\mathbf{X} \notin \mathcal{B}_2$, we add the equations

$$p_{\mathbf{X}} = p_{(\text{AC})\mathbf{X}} \quad p_{\mathbf{X}} = p_{(\text{AT})\mathbf{X}}.$$

Each orbit provides a couple of equations. To avoid repetitions, we choose \mathbf{X} of the form $(\mathbf{A}^l)(\mathbf{C}^m)\mathbf{x}_{l+m+1} \dots \mathbf{x}_n$ (where $l, m \geq 1$) and such that if T appears in \mathbf{X} , there is some \mathbf{G} in a preceding position. The number of such equations is $(3 - 1) \times \left(\frac{2^{2n-3}+1}{3} - 2^{n-2}\right) = \frac{2^{2n-2}+2}{3} - 2^{n-1}$.

□

Remark 4.12. A rather different approach based on representation theory can be considered to compute the equations for $\mathcal{L}^{G\mathcal{M}}$. We explain the idea roughly: for each group G under consideration, take the decomposition of \mathcal{L} in isotypic components induced by the defining representation of G :

$$(7) \quad \mathcal{L} = \bigoplus_{i=1}^s \mathcal{L}[\omega_i]$$

Then, $\mathcal{L}^{G\mathcal{M}}$ is just the component corresponding to the trivial representation, which maps every element $g \in G$ to the identity map on \mathcal{L} . The maps $\theta_{\omega_i} = \frac{1}{|G|} \sum_{g \in G} \omega_i(g)g$ define projections

$$\theta_{\omega_i} : \mathcal{L} \rightarrow \mathcal{L}[\omega_i].$$

Since the isotypic components are orthogonal, we can proceed to systematically apply these projections to obtain basis for the non-trivial isotypic components. The inner product $\langle \cdot, \cdot \rangle$ defined in (5) can then be used to infer a minimal system of equations defining $\mathcal{L}^{G\mathcal{M}}$ from these basis.

Remark 4.13. The sets of equations provided in Theorem 4.6 has been successfully used in the paper [KDGC11] for model selection. Although the dimensions of these linear spaces are exponential in n , for its biological application one does not need to consider all the equations but only those containing the patterns observed in the data (in real applications the number of different columns in an alignment is really small compared to the dimension of these spaces). Algorithm

SPI_n implemented for the purpose of the above paper selects the equations in Theorem 4.6 in this way. Moreover, as the equations provided in this theorem are binomials, these equations are useful for obtaining the maximum likelihood estimate and applying Akaike's information criterion (see [KDGC11] for details).

Example 4.14. 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}^{SSM} : \mathbb{E}^{SSM} is composed of the following equations:

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

Equations for \mathcal{L}^{K81} : \mathbb{E}^{K81} is composed of \mathbb{E}^{SSM} together with

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

Equations for \mathcal{L}^{K80} : \mathbb{E}^{K80} is composed of \mathbb{E}^{K81} together with

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

Equations for $\mathcal{L}^{\text{JC69}}$: \mathbb{E}^{JC69} is composed of \mathbb{E}^{K80} together with

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

5. IDENTIFIABILITY OF PHYLOGENETIC MIXTURES

Definition 5.1. 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 the details of this definition). Similarly, we can define the join of projective varieties $X_1, \dots, X_h \subset \mathbb{P}^m$, $\bigvee_{i=1}^h X_i$, 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

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

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

For example, 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

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

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

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

where $\Omega = \{\mathbf{a} = (a_1, \dots, a_h) \mid \sum_i a_i = 1\}$ 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.

We recall the definition of generic identifiability of the tree topologies on h -mixtures (see for example [APRS10]).

Definition 5.2. The *tree topologies* on h -mixtures over \mathcal{M} are *generically identifiable* if for any set of trivalent tree topologies T_1, \dots, T_h and generic choice of $(\xi_1, \dots, \xi_h, \mathbf{a}) \in Par_{s\mathcal{M}}(T_1) \times \dots \times Par_{s\mathcal{M}}(T_h) \times \Omega$, 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 stochastic parameters $(\xi'_1, \dots, \xi'_h, \mathbf{a}')$, 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 viceversa.

The tree topologies are the discrete parameters of h -mixtures. When we come to the continuous parameters we have the following definition.

Definition 5.3. The *continuous parameters* on h -mixtures on T_1, \dots, T_h under an evolutionary model \mathcal{M} are *generically identifiable* if for generic choices 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 $(\xi_1, \dots, \xi_h, \mathbf{a}) = (\xi'_1, \dots, \xi'_h, \mathbf{a}')$ or an allowed permutation of the parameters (see [APRS10, Definition 2]).

In terms of algebraic varieties, generic identifiability of continuous parameters implies that the generic fibers of the map $\phi_{T_1} \vee \dots \vee \phi_{T_h}$ are finite. In particular, the fiber dimension theorem applies (cf. [Har92, Theorem 11.12]) to obtain

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

The converse of this result (that is, finite generic fibers of $\phi_{T_1} \vee \dots \vee \phi_{T_h}$ imply generic identifiability) is not necessarily true because a finite fiber can be formed by more than one point stochastically meaningful.

Example 5.4. The tree topologies and the continuous parameters are generically identifiable for the unmixed equivariant models JC69, K80, K81, SSM, GMM (see [CFS11, Corollary 3.9]).

If the continuous parameters are generically identifiable under an evolutionary model \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 and corresponds to the number of free parameters of the stochastic model (fiber dimension theorem cf. [Har92, Theorem 11.12]). Let $d_{\mathcal{M}}$ be this dimension, then we have the following result.

Theorem 5.5. *Let \mathcal{M} be an evolutionary model for which continuous parameters are generically identifiable on trivalent trees and let $h_0 := \frac{\dim \mathcal{D}_{\mathcal{M}}}{d_{\mathcal{M}}+1}$ where $d_{\mathcal{M}}$ is the dimension of $\mathbb{P}V_T^{\mathcal{M}}$ as above. Then either the continuous parameters or the tree parameters are not generically identifiable for h -mixtures under the model \mathcal{M} if $h \geq h_0$.*

Remark 5.6. This theorem proves that it makes no sense to do phylogenetic inference for h -mixtures when $h \geq h_0$.

Corollary 5.7. *Let $[n]$ be a set of taxa and \mathcal{M} be one of the equivariant models JC69, K80, K81, SSM, GMM. Then phylogenetic h -mixtures under these models are not identifiable for $h \geq h_0$ where*

- $h_0 = \frac{4^n}{12(2n-3)+4}$ if $\mathcal{M} = \text{GMM}$,

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

Proof. Theorem 4.2 shows that $\mathcal{L}^{\mathcal{M}} = \mathcal{D}_{\mathcal{M}}$ and Proposition 4.4 gives the dimension of this space in each case. Then, we apply Theorem 5.5 taking into account that $d_{\text{GMM}} = 12(2n-3) + 3$, $d_{\text{SSM}} = 6(2n-3) + 1$, $d_{\text{K81}} = 3(2n-3)$, $d_{\text{K80}} = 2(2n-3)$ and $d_{\text{JC69}} = 2n-3$. \square

Example 5.8. Consider the Kimura 3-parameter model K81 and consider trees on $n = 4$ taxa. Then for any $h \geq 4$, phylogenetic h -mixtures are not identifiable (Corollary 5.7). We are not aware of any result proving that mixtures of 2 or 3 different tree topologies under this model are identifiable (either for tree parameters or for continuous parameters).

Example 5.9. If we consider the Jukes-Cantor model JC69 on $n = 4$ taxa, then Corollary 5.7 tells us that for $h \geq 3$, h -mixtures are not identifiable. Therefore for this particular model on four taxa the identifiability is solved: the tree and continuous parameters are generically identifiable for the unmixed model; the tree parameters are generically identifiable for 2-mixtures [APRS10, Theorem 10]; the continuous parameters are generically identifiable for 2-mixtures on different tree topologies and not identifiable for the same tree topology [APRS10, Theorem 23]; either the continuous parameters or the tree topologies are not generically identifiable for more than two mixtures (Corollary 5.7).

Proof of Theorem 5.5. Let $\text{edim}(h) := hd_{\mathcal{M}} + h - 1$. Then the variety $\vee_{i=1}^h \mathbb{P}V_{T_i}$ has dimension $\leq \text{edim}(h)$. Indeed, as $\vee_i \phi_{T_i}$ is a parameterization of an open subset of $\vee_{i=1}^h \mathbb{P}V_{T_i}$, then the dimension of $\vee_{i=1}^h \mathbb{P}V_{T_i}$ is less or equal than $\sum \dim \mathbb{P}V_{T_i} + h - 1$. Moreover, the dimension of $\mathbb{P}V_{T_i}$ is equal to $d_{\mathcal{M}}$ if T_i is trivalent (because the continuous parameters for the unmixed models we are considering are generically identifiable) and is less than $d_{\mathcal{M}}$ for non-trivalent trees. Therefore $\dim(\vee_{i=1}^h \mathbb{P}V_{T_i}) \leq \text{edim}(h)$.

If we consider only trivalent trees T_i , 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 fiber dimension theorem applied to $\vee \phi_{T_i}$, equality 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.

If $h_0 = \frac{\dim \mathcal{D}_{\mathcal{M}}}{d_{\mathcal{M}}+1}$ then, $\text{edim}(h_0) = h_0(d_{\mathcal{M}} + 1) - 1 = \dim \mathcal{D}_{\mathcal{M}} - 1$. Now we fix an $h \in \mathbb{N}$ with $h \geq h_0$, so that one has $\text{edim}(h) \geq \dim(\mathcal{D}_{\mathcal{M}}) - 1$.

Two things could happen:

(a) For all tree topologies $\{T_1, \dots, T_h\}$ one has $\dim(\bigvee_{i=1}^h \mathbb{P}V_{T_i}) < \dim(\mathcal{D}_{\mathcal{M}}) - 1$.

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

Case (a) implies that for any set of trivalent tree topologies $\{T_1, \dots, T_h\}$ one has $\dim(\bigvee_{i=1}^h \mathbb{P}V_{T_i}) < \text{edim}(h)$. And we have seen above that this implies that the continuous parameters are not generically identifiable.

In case (b) one has that $\bigvee_{i=1}^h \mathbb{P}V_{T_i} = \mathbb{P}(\mathcal{D}_{\mathcal{M}})$. Indeed, $\bigvee_{i=1}^h \mathbb{P}V_{T_i} \subset \mathbb{P}(\mathcal{D}_{\mathcal{M}})$ and $\dim(\bigvee_{i=1}^h \mathbb{P}V_{T_i}) = \dim(\mathcal{D}_{\mathcal{M}}) - 1 = \dim(\mathbb{P}(\mathcal{D}_{\mathcal{M}}))$ which implies that both varieties coincide (the proper subvarieties of an affine space have dimension strictly smaller than it). In particular any other h -mixture (which is a point in $\mathbb{P}(\mathcal{D}_{\mathcal{M}})$) would be contained in $\bigvee_{i=1}^h \mathbb{P}V_{T_i}$ and therefore the topologies are not generically identifiable. \square

Remark 5.10. The negative result of Theorem 5.5 should be complemented with the following positive result of Rhodes and Sullivant in [RS]: if $\mathcal{M} = \mathbf{GMM}$ 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 $h < 4^{\lceil \frac{n}{4} \rceil - 1}$.

REFERENCES

- [APRS10] ES Allman, S Petrovic, JA Rhodes, and S Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 2010. to appear.
- [AR04] ES Allman and JA Rhodes. Quartets and parameter recovery for the general Markov model of sequence mutation. *AMRX 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. Computational Algebraic Statistics.
- [AR08] ES Allman and JA Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Adv. in Appl. Math.*, 40:127–148, 2008.
- [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. *Syst Biol.*, 2011.
- [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] YX 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.
- [KDGC11] A Kedzierska, M Drton, R Guigó, and M Casanellas. SPIn: model selection for phylogenetic mixtures via linear invariants. To appear in *Molecular Biology and Evolution*, 2011.
- [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.
- [RS] JA Rhodes and S Sullivant. Identifiability of large phylogenetic mixture models. <http://arxiv.org/abs/1011.4134v1>.
- [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.
- [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(6):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. *J. Comput. Biol.*, 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.

E-mail address: `jesus.fernandez.sanchez@upc.edu`

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

E-mail address: `anna.kedzierska@upc.edu`