Relevant phylogenetic invariants of evolutionary models

Marta Casanellas[a,1], Jesús Fernández-Sánchez[a,1]

[a] *Dpt. Matemàtica Aplicada I. ETSEIB-UPC. Avinguda Diagonal 647. 08028 Barcelona. Spain*

Abstract

(English) Recently there have been several attempts to provide a whole set of generators of the ideal of the algebraic variety associated to a phylogenetic tree evolving under an algebraic model. These algebraic varieties have been proven to be useful in phylogenetics. In this paper we prove that, for phylogenetic reconstruction purposes, it is enough to consider generators coming from the edges of the tree, the so-called edge invariants. This is the algebraic analogous to Buneman's Splits Equivalence Theorem. The interest of this result relies on its potential applications in phylogenetics for the widely used evolutionary models such as Jukes-Cantor, Kimura 2 and 3 parameters, and General Markov models.

(French) Dans les dernières années, il y a eu différentes tentatives pour apporter un ensemble complet de générateurs de l'idéal d'une variété algébrique associée à un arbre phylogénétique qui évolue sous un modèle algébrique. Ces variétés algébriques ont montré leur utilité en phylogénétique. Dans cet article, on prouve que, pour la reconstruction phylogénétique, il est suffisant de considérer certains générateurs obtenus des arêtes de l'arbre, qu'on appelle invariants des arêtes. C'est l' équivalent algébrique du Théorème de Buneman au sujet de l'équivalence des bipartitions induites par les arêtes. L'intérêt de ce résultat se base en ses applications potentielles en phylogénétique pour les modèles les plus utilisées, comme le modèle de Jukes-Cantor, le modèle de Kimura avec 2 ou 3 paramètres, et le modèle général de Markov.

Keywords: phylogenetic invariants, phylogenetic varieties, molecular evolutionary models
2000 MSC: 14J99, 92D15, 05C85

[1] Both authors are partially supported by Ministerio de Educación y Ciencia, MTM2009-14163-C02-02, and Generalitat de Catalunya, 2009 SGR 1284.

1. Introduction

Algebraic evolutionary models and the algebraic varieties associated to a tree evolving under these models have been an interdisciplinary area of research with successful results in the last five years. The use of polynomials in phylogenetic reconstruction was first introduced by biologists Cavender and Felsenstein [11] and Lake [20]. Because of their interest in phylogenetics, there have been several attempts to provide a set of generators of the ideal of these algebraic varieties (see for example [3], [25], [12], [10]). On the other hand, the authors of this paper have proven in [7] that these generators can be successfully used in phylogenetic reconstruction. In other words, methods based in algebraic geometry can lead to the inference of the phylogenetic tree of current biological species. As we already did in [8], our aim in the present paper is to address again the study of these algebraic varieties towards their real applications in phylogenetics.

Algebraic evolutionary models include the algebraic version of widely used models in biology such as Jukes-Cantor model [17], Kimura 2 and 3 parameters model (cf. [18], [19]) and the general Markov model (cf. [5]). These models belong to what Draisma and Kuttler call *equivariant models* in [12] (see section 2 for the precise definition). Following ideas of Allman and Rhodes and using representation theory, Draisma and Kuttler have recently given an algorithm to obtain the generators of the ideal of the algebraic varieties associated to a tree of n species evolving under an equivariant model from the generators of the ideal associated to a tree of 3 species and certain minors of matrices (the so-called *edge invariants*). Nevertheless, a set of generators for trees of 3 species is not known for certain models such as the general Markov model (this is the so-called Salmon Conjecture) or the strand symmetric model (see [10]). Therefore, a complete list of generators for a tree of n species evolving under these models cannot be given at this point.

The goal of this paper is to prove that, whereas mathematically speaking it is interesting to know a set of generators of the ideal of these varieties, for biological purposes it is enough to consider certain generators. More precisely, the edge invariants mentioned above suffice to reconstruct the phylogenetic tree of any number of species (see the Theorem in the next page or Theorem 4.4). This is a natural result if one thinks of the combinatorics result of Buneman that says that a tree can be recovered if one knows the set of splits on the set of leaves induced by its edges (cf. [6], [21, Theorem 2.35], see also Theorem 4.1 below).

Our inspiration goes back to the work [14] of biologist Joe Felsenstein who calls *phylogenetic invariants* those polynomial expressions that vanish on the expected frequencies of any sequences arising from one tree topology but are non zero for at least one tree of another topology. A tree topology in this setting is the topology of the tree graph labelled at the leaves with the name of the species. Algebraically speaking, he calls *phylogenetic invariants* those elements of the ideal associated to a phylogenetic tree that allow to distinguish it from other tree topologies. In the mathematical context, the name phylogenetic invariants has usually been given to all elements of the ideal, see for instance the work of Allman and Rhodes [3]. We want to go back to the original meaning of phylogenetic invariants because our focus is devoted to the applications of algebraic geometry in the reconstruction of the tree topology of current species. Therefore, we are mainly interested in precisely those elements of the ideal that provide information for phylogenetic reconstruction purposes; in other words, we are interested in *phylogenetic* invariants (i.e polynomials in the ideal of one tree topology of n species but not in the ideal of all other tree topologies on the same number of species) and the word *invariants* alone shall mean any element of the ideal. In colloquial language the main result of this paper is that, for phylogenetic reconstruction purposes, the relevant phylogenetic invariants are the edge invariants mentioned above.

As our aim is to study these varieties regarding their applications in biology, let us roughly explain here how does algebraic geometry interfere with phylogenetic reconstruction. Let n be a number of biological species and assume that we are given an alignment of DNA sequences corresponding to them (the definition of alignment is rather technical but it refers to a collection of n-tuples in $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^n$ that will be also called columns of the alignment). Each column stands for sites in the n DNA sequences that have evolved from the same nucleotide in the common ancestor. We assume that these species are leaves of a phylogenetic tree T evolving under a probabilistic model $\mathcal{M}$ (in this paper we will only consider equivariant models, see Definition 2.4 for the precise definition). It is usual to assume as well that all columns of the alignment behave independently and identically (i.e. all sites of the DNA sequences of these species evolve in the same way and independently of the other sites). Associated to this model $\mathcal{M}$ there is a parameterization map Ψ_T giving the joint distribution of states $\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}$ at the leaves of T as polynomial functions of continuous parameters. Therefore, as an alignment of DNA sequences evolving under this model on a tree T is a collection of observations of states at the leaves, it corresponds to a point in the image of this parameterization map. The algebraic variety $V_{\mathcal{M}}(T)$ associated to T is the closure of this image (see Definition 2.7). In the real life, alignments are not points of $V_{\mathcal{M}}(T)$ but they are *close* to $V_{\mathcal{M}}(T)$ if the model reasonably fits the data. Therefore the idea behind phylogenetic algebraic geometry is to use the ideal of $V_{\mathcal{M}}(T)$ in order to infer the tree topology T. See [9] for an algorithm of phylogenetic reconstruction based on the generators of this ideal and [7] for tests of it on simulated data.

Up to now, all attempts have focused on giving a whole set of generators of $I(V_{\mathcal{M}}(T))$ but our approach is more practical. As biologists assume that the model $\mathcal{M}$ fits the data, the point given by an alignment is therefore assumed to be close to the union of all varieties $V_{\mathcal{M}}(T)$ for trees of n species evolving under model $\mathcal{M}$. Henceforth, we only need to know how a particular variety $V_{\mathcal{M}}(T_0)$ is defined inside $\cup_T V_{\mathcal{M}}(T)$ where the union runs over all trivalent tree topologies T of n species. In this algebraic geometry context our main result (Theorem 4.4) can be summarized in the following way.

Theorem. *Let $\mathcal{T}$ be the set of trivalent tree topologies on n leaves and let $\mathcal{M}$ be an equivariant model. For each tree topology $T \in \mathcal{T}$ there exists an open set U_T such that if p belongs to $\cup_{T\in\mathcal{T}} U_T$, then p belongs to a particular variety $V_{\mathcal{M}}(T_0)$ if and only if p belongs to the zero set of the edge invariants of T_0.*

This result has also other consequences in phylogenetics. For instance, it says that edge invariants should not be used for model fitting tests (see [16] for an algebraic introduction to the subject) or for the study of identifiability of continuous parameters (see [4] for an explanation of these terminology) of the model because they are indeed phylogenetic invariants. Instead, they should be used in discussing the identifiability of tree topology of such models (see Corollary 3.10) as it was already done by Allman and Rhodes in [2]. We also find invariants (not phylogenetic invariants) that could potentially be used for model fitting tests, that is, linear polynomials that can be used for choosing the evolutionary model that best fits the data (see Remark 2.7).

Moreover, our main theorem allows one to give the exact degrees of those generators relevant in phylogenetics (see Corollary 4.12), whereas the degrees of a whole set of generators for the general Markov or strand symmetric models are still unknown. It is worth highlighting that these degrees can be computed by just knowing the model we are interested in, and they do not depend on the topology or the number of leaves we are considering.

Here we outline the structure of the paper. In section 2 we adapt the setting and notation of [12] to our convenience. As well, we prove and recall basic facts of group representation theory for those non-familiarized readers. Section 3 is devoted to prove a technical result that will be the key in the proof of our main theorem. Roughly speaking this result proves that edge invariants are indeed *phylogenetic* invariants for any equivariant model. This was already known for the general Markov model by Allman, Rhodes (see for instance [2]) and Eriksson [13] but it is new for the remaining equivariant models. The proof relies on providing a formula for the rank of the flattening of the tensor Ψ_T along any bipartition of the set of leaves. In section 4 we prove Theorem 4.4, our main result. In the last section we provide an exhaustive collection of examples on how to compute the required edge invariants for the most used evolutionary models: Jukes-Cantor, Kimura 2 and 3 parameters, strand symmetric and general Markov model. We compute them explicitly for quartet trees. It is our aim to make this section clear enough for biomathematicians so that, for example, we relate invariants used by biologist like Lake (see [20]) to the more technical definition of edge invariants (see the end of subsection 5.5). We also connect our edge invariants to Fourier coordinates that are more familiar to those readers used to group-based models. In particular, the reader can visualize what are the Fourier coordinates that are actually interesting in biology as not all of them are needed for phylogenetic reconstruction. This section is also a useful illustration of technical definitions given in sections 2 and 3 so it is a good idea to combine the reading of both sections with section 5.

Acknowledgments: We would like to thank Josep Elgueta and Jeremy Sumner for useful comments on group representation theory.

2. Preliminaries

A *tree* is a connected finite graph without cycles, consisting of vertices and edges. Given a tree T, we write $V(T)$ and $E(T)$ for the set of vertices and edges of T. The *degree* of a vertex is the number of edges incident on it. The set $V(T)$ splits into the set of *leaves* $L(T)$ (vertices of degree one) and the set of interior vertices $\mathrm{Int}(T)$: $V(T) = L(T) \cup \mathrm{Int}(T)$. One says that a tree is *trivalent* if each vertex in $\mathrm{Int}(T)$ has degree 3. A *tree topology* is the topological class of a tree where every leaf has been labelled. Given a subset L of $L(T)$, the *subtree induced by* L is just the smallest tree composed of the edges and vertices of T in any path connecting two leaves in L.

Given an ordered set $B = \{b_1, b_2, \dots, b_k\}$, we define $W = \langle B \rangle_{\mathbb{C}}$ as the $\mathbb{C}$-vector space generated by the elements of B. For biological applications, the most common values of k are 2, 4 or 20 (for example, $B = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$). Now, given a subgroup G of the group $\mathfrak{S}_k$ of permutations of k elements, we consider the restriction to G of the natural linear representation

$$\rho : \mathfrak{S}_k \to GL(W)$$

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

$$g \cdot u := \rho(g)(u) \in W.$$

In fact, ρ induces a G-module structure on any tensor power of W, say $\otimes^l W := W \otimes \ldots \otimes W$, by taking

$$g \cdot (u_1 \otimes \ldots \otimes u_l) := g \cdot u_1 \otimes \ldots \otimes g \cdot u_l. \tag{2.1}$$

Henceforth, any tensor power of W will be implicitly considered as a G-module with this action.

From now on, we fix an ordered set $B = \{b_1, b_2, \ldots, b_k\}$, $W = \langle B \rangle_{\mathbb{C}}$ and a subgroup $G \subset \mathfrak{S}_k$ acting on W as above.

Definition 2.1. A *phylogenetic tree on* (G, W) is a tree where every vertex p has a $\mathbb{C}$-vector space $W_p \cong W$ associated to it, regarded as a representation of G via the map ρ defined above.

Notation. The scalar product with orthonormal basis B_p will be denoted by $(. \mid .)_p$. This gives a canonical isomorphism from W_p to W_p^*.

Notice that the scalar product $(. \mid .)_p$ is G-invariant, that is, $(g{\cdot}u \mid g{\cdot}v)_p = (u \mid v)_p$ for every $u, v \in W_p$ and any $g \in G$.

Definition 2.2. Given a phylogenetic tree T on (G, W), a T-*tensor* is any element of

$$\mathcal{L}(T) := \otimes_{p \in L(T)} W_p.$$

A G-*tensor* on T is a T-tensor invariant by the action defined in (2.1). The set of G-tensors will be denoted by $\mathcal{L}(T)^G$.

From now on, if $l > 0$ we write $\otimes^l W = W \otimes \overset{l}{\ldots} \otimes W$. We denote by $B(\otimes^l W)$ the basis of $\otimes^l W$ given by

$$\{u_{i_1} \otimes \ldots \otimes u_{i_l} \mid u_{i_j} \in B\}.$$

This is an othonormal basis with respect to the scalar product of $\otimes^l W$ given by $(\otimes_p u_p \mid \otimes_p v_p) = \prod_p (u_p \mid v_p)$. If $L \subset L(T)$ is a subset of $L(T)$ and $l = \sharp L$, then we shall use the notation $\otimes_L W$ for the space $\otimes_{p \in L} W_p \cong \otimes^l W$.

Definition 2.3. Let T be a phylogenetic tree on (G, W) and assume that a distinguished vertex of T (the *root*) is given, inducing an orientation in all the edges of T: write e_0 and e_1 for the origin and final vertices of the edge e, respectively. A G-*evolutionary presentation*[2] of T is a collection of tensors $\{A_{e_0,e_1}\}_{e \in E(T)}$ where each

[2] Notice that *evolutionary presentations* are called *representations* in [12]. We prefer this terminology to avoid confusion with representation theory.

A_{e_0,e_1} is a G-invariant element of the G-module $W_{e_0} \otimes W_{e_1}$. The space of G-invariant elements of $W_{e_0} \otimes W_{e_1}$ is denoted by $(W_{e_0} \otimes W_{e_1})^G$.

If another root (orientation) on T is considered, inducing the opposite orientation on some edge $e \in E(T)$, we define $A_{e_1,e_0} := A^t_{e_0,e_1}$, where $.^t$ is the natural isomorphism $(W_{e_0} \otimes W_{e_1})^G \cong (W_{e_1} \otimes W_{e_0})^G$. We will often identify $\mathrm{Hom}_G(W_{e_0}, W_{e_1})$ with $(W_{e_0} \otimes W_{e_1})^G$ via $W^*_{e_0} \cong W_{e_0}$. With this convention, G-evolutionary presentations on a tree do not depend on the orientation chosen. The space of all G-evolutionary presentations of T is the parameter space denoted by $\mathrm{Par}_G(T) = \prod_{e \in E(T)} (W_{e_0} \otimes W_{e_1})^G$. Notice that a G-evolutionary presentation of T induces by restriction a G-evolutionary presentation of any subtree of T.

The space $\mathrm{Par}_G(T)$, as well as $\mathcal{L}(T)$ and $\mathcal{L}(T)^G$, are irreducible affine spaces with their Zariski topology.

Definition 2.4. An *equivariant model* of evolution is a pair (G, W) as above, $W = \langle b_1, \ldots, b_k \rangle$, $G \subset \mathfrak{S}_k$. Trees evolving under this equivariant model are phylogenetic trees on (G, W) together with the space of G-evolutionary presentations.

Equivariant models of evolution include the general Markov model [5] when $G = \{\mathrm{id}\}$, the strand symmetric model [10] when $G = \langle (\mathtt{AT})(\mathtt{CG}) \rangle$, and the algebraic versions of Kimura 3-parameters [19] ($G = \langle (\mathtt{AC})(\mathtt{GT}), (\mathtt{AG})(\mathtt{CT}) \rangle$), Kimura 2-parameters [18] ($G = \langle (\mathtt{ACGT}), (\mathtt{AG}) \rangle$) and Jukes-Cantor models [17] ($G = \mathfrak{S}_4$). We derive the reader to section 5 for specific computations with these models.

Following [3] and [12] we present now a fundamental operation $*$ on phylogenetic trees, G-evolutionary presentations and T-tensors. To this aim, we first introduce a bilinear operation $\langle \cdot \mid \cdot \rangle$ between tensors induced by the bilinear form $(\cdot \mid \cdot)$ on W. Let X and Y be two finite sets of indices with $Z = X \cap Y \neq \emptyset$, and such that every p in X or Y has associated a vector space $W_p \cong W$ to it. Define the contraction map as

$$\begin{array}{rcl} \langle . \mid . \rangle : \otimes_X W \times \otimes_Y W & \rightarrow & \otimes_{X \cup Y \setminus Z} W \\ (\otimes_{p \in X} v_p, \otimes_{p \in Y} u_p) & \mapsto & (\otimes_{p \in Z} v_p \mid \otimes_{p \in Z} u_p) \big((\otimes_{p \in X \setminus Z} v_p) \otimes (\otimes_{p \in Y \setminus Z} u_p) \big) \end{array} \quad (2.2)$$

Now, we define the $*$ operation:

$*$ **for trees:** Given l phylogentic trees $T_1, \ldots, T_l$ on (G, W) whose vertex sets only share a common leaf q with common space W_q and common basis B_q, we construct a new tree $*_i T_i$ on (G, W) obtained by gluing the T_i's' along q; the space at a vertex of $*_i T_i$ coming from T_j is just the space attached to it in T_j, with the same distinguished basis.

$*$ **for G-evolutionary presentations:** Given G-evolutionary presentations $A_i \in \mathrm{Par}_G(T_i)$ for $i = 1, ..., l$, we denote by $*_i A_i$ the G-evolutionary presentation of $*_i T_i$ built up from the A_i.

$*$ **for tensors:** Now let ψ_i be a T_i-tensor, for all i. Then we obtain a T-tensor as follows:

$$*_i \psi_i := \sum_{b \in B_q} \otimes_i \langle b \mid \psi_i \rangle.$$

Although this $*$ operator is not a binary operator extended to several factors, when convenient we will write $T_1 * \ldots * T_l$ for $*_i T_i$ and $\psi_1 * \ldots * \psi_l$ for $*_i \psi_i$.

Now we describe a basic procedure that allows us to associate a T-tensor to any G-evolutionary presentation of T. We proceed inductively on the number of edges to define $\Psi_T : \mathrm{Par}_G(T) \to \mathcal{L}(T)$. Let $A \in \mathrm{Par}_G(T)$. First, if T has a single edge p, q, then $\Psi_T(A) := A_{qp}$, is an element of $\mathcal{L}(T) = W_q \otimes W_p$. If T has more than one edge, then let q be any internal vertex of T. Two vertices $p, q \in T$ are *adjacent* if they are joined by an edge; in this case, we write $p \sim q$. We can then write $T = *_{p \sim q} T_p$, where T_p is the branch of T around q containing p, constructed by taking the connected component of $T \setminus \{q\}$ containing p, and reattaching q to p. The G-evolutionary presentation A induces G-evolutionary presentations A_p of the T_p, and by induction $\Psi_{T_p}(A_p)$ has been defined. We now set

$$\Psi_T(A) := *_{p \sim q} \Psi_{T_p}(A_p).$$

This definition is independent of the choice of q and the formula is also valid if q is actually a leaf (see [12] for details). Moreover, we have that the map $\Psi_T : \mathrm{Par}_G(T) \to \mathcal{L}(T)$ is G-equivariant (see [12, Lemma 5.1]), so that $\mathrm{Im}\Psi_T \subset \mathcal{L}(T)^G$.

Remark 2.5. Notice that the above map $\Psi_T : \mathrm{Par}_G(T) \to \mathcal{L}(T)^G$ is a continuous map in the Zariski topology.

Definition 2.6. The *algebraic variety associated to a phylogenetic tree* T on (G, W) is

$$V_G(T) := \overline{\{\Psi_T(A) \mid A \in \mathrm{Par}_G(T)\}} \subset \mathcal{L}(T)$$

where the closure is taken in the Zariski topology.

Notice that we have $V_G(T) \subset \mathcal{L}(T)^G$. From now on, we will consider $\mathcal{L}(T)^G$ as the ambient space of $V_G(T)$ and $\mathcal{I}(T)$ will be the ideal of this variety in the corresponding coordinate ring. When the group is understood from the context, we will use the notation $V(T)$.

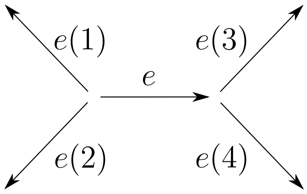


Figure 1:

Remark 2.7. The inclusion $\mathcal{L}(T)^G \subseteq \mathcal{L}(T)$ is defined by a set of linear polynomials that are also *invariants* of any phylogenetic tree T on (G, W) (see the Introduction for the explanation of the word invariants). Although they are not *phylogenetic invariants* because they vanish on $V_G(T)$ for any tree T, they might be interesting for choosing the model (G, W) that best fits the data. This application of invariants to model fitting will be studied in a forthcoming paper.

Example 2.8. If we consider $B = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and $G = \{\mathrm{id}\} \subset \mathfrak{S}_4$, we obtain the general Markov model. In this case, $(W_{e_0} \otimes W_{e_1})^G = (W_{e_0} \otimes W_{e_1})$ and no restrictive conditions are imposed on the parameters of the model. Thus, a G-evolutionary presentation can be identified, by taking the basis B in W with a collection of matrices $\{A_e\}_{e \in E(T)}$ and the parameters of the model are the entries of these matrices. When these entries are real non-negative values and their columns sum to 1, they can be understood as the probabilities of substitution among the 4 nucleotides:

$$A_e = \begin{pmatrix} P(\mathtt{A} \mid \mathtt{A}, e) & P(\mathtt{A} \mid \mathtt{C}, e) & P(\mathtt{A} \mid \mathtt{G}, e) & P(\mathtt{A} \mid \mathtt{T}, e) \\ P(\mathtt{C} \mid \mathtt{A}, e) & P(\mathtt{C} \mid \mathtt{C}, e) & P(\mathtt{C} \mid \mathtt{G}, e) & P(\mathtt{C} \mid \mathtt{T}, e) \\ P(\mathtt{G} \mid \mathtt{A}, e) & P(\mathtt{G} \mid \mathtt{C}, e) & P(\mathtt{G} \mid \mathtt{G}, e) & P(\mathtt{G} \mid \mathtt{T}, e) \\ P(\mathtt{T} \mid \mathtt{A}, e) & P(\mathtt{T} \mid \mathtt{C}, e) & P(\mathtt{T} \mid \mathtt{G}, e) & P(\mathtt{T} \mid \mathtt{T}, e) \end{pmatrix}.$$

Here $P(\mathtt{X} \mid \mathtt{Y}, e)$ is the conditional probability that nucleotide $\mathtt{Y}$ at the parent species e_0 is being substituted along edge e by nucleotide $\mathtt{X}$ at its child species e_1. In our terminology introduced above, $P(\mathtt{X} \mid \mathtt{Y}, e)$ is the coordinate of $A_e \in W_{e_0} \otimes W_{e_1} \cong W \otimes W$ corresponding to $\mathtt{Y} \otimes \mathtt{X}$. Given a tree T, the G-equivariant map Ψ_T is the parameterization that associates to each parameter set the vector of expected pattern frequencies $p = (p_{\mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_n})_{\mathtt{X}_i \in B}$ (that is, $p_{\mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_n}$ is the probability of observing $\mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_n$ at the leaves of T). For example, if T is a 4-leaf tree as in figure 1, then

$$\begin{array}{rcl} \Psi_T : \prod_{e \in E(T)} (W \otimes W) \cong \mathbb{C}^{80} & \to & \otimes^4 W \cong \mathbb{C}^{256} \\ (A_e)_e & \mapsto & (p_{\mathtt{AAAA}}, p_{\mathtt{AAAC}}, \ldots, p_{\mathtt{TTTT}}) \end{array}$$

and $p_{\mathtt{X}_1\mathtt{X}_2\mathtt{X}_3\mathtt{X}_4}$ is the coordinate of $p \in \mathcal{L}(T) \cong \mathbb{C}^{256}$ corresponding to the basis vector $\mathtt{X}_1 \otimes \mathtt{X}_2 \otimes \mathtt{X}_3 \otimes \mathtt{X}_4$. In this case, the image of Ψ_T is given by

$$p_{\mathtt{X}_1\mathtt{X}_2\mathtt{X}_3\mathtt{X}_4} = \sum_{\mathtt{Y},\mathtt{Z}} \pi_{\mathtt{Y}} A_e(\mathtt{Z},\mathtt{Y}) A_{e(1)}(\mathtt{X}_1,\mathtt{Y}) A_{e(2)}(\mathtt{X}_2,\mathtt{Y}) A_{e(3)}(\mathtt{X}_3,\mathtt{Z}) A_{e(4)}(\mathtt{X}_4,\mathtt{Z}).$$

Here $\pi_{\mathtt{Y}}$ is the probability of nucleotide $\mathtt{Y}$ occurring at the root node (see figure 1). Actually, in the original definition of Ψ_T (see paragraph before Remark 2.5) we gave a reparameterization of $V_G(T)$ where we omit parameters $\pi_{\mathtt{Y}}$ for convenience.

Definition 2.9. Given a tree T, a *bipartition* of the leaves of T is a decomposition $L(T) = L_1 \cup L_2$ where $L_1 \cap L_2 = \emptyset$. We denote it as $L_1 \mid L_2$. We say that $L_1 \mid L_2$ is non-trivial bipartition if $\sharp L_1 \geq 1$ and $\sharp L_2 \geq 1$. Notice that every edge e of T induces a non-trivial bipartition $L_1 \mid L_2$ of $L(T)$ by removing it; such a bipartition is called an *edge split* of T and will be denoted by the same letter e.

2.1. Representation Theory

We will make use of representation theory of groups. A basic reference for this are the books [22] and [15] and the reader is referred to them for definitions and well-known facts.

From now on, write $\Omega_G = \{\omega_1, \ldots, \omega_s\}$ for the set of irreducible characters of G.

It is known that any two representations with the same character are isomorphic (Corollary 2 of § 2 of [22]). As a consequence of this and Schur's lemma (see §2.2 of [22]) we obtain the following fundamental result in representation theory:

Lemma 2.10. *Let $N_\omega, N_{\omega'}$ be the irreducible linear representations of G with associated characters $\omega, \omega' \in \Omega_G$. If $f : N_\omega \to N_{\omega'}$ is a G-module homomorphism, and*

(i) if $\omega \neq \omega'$, then $f = 0$;

(ii) if $\omega = \omega'$, then f is a homothety.

In particular, $\mathrm{Hom}_G(N_\omega, N_\omega) \cong \mathbb{C}$.

For every irreducible character $\omega_t \in \Omega_G$, fix an irreducible G-module N_{ω_t} with associated character ω_t. Then, for any G-module V, there exists a unique decomposition of V into isotypic components:

$$V \cong \oplus_{t=1}^{s} V[\omega_t] \tag{2.3}$$

where each $V[\omega_t]$ is isomorphic to $N_{\omega_t} \otimes \mathbb{C}^{m(\omega_t,V)}$ for some multiplicity $m(\omega_t, V)$, $t = 1, \dots, s$. We also have that if V' is another representation of G, then

$$\mathrm{Hom}_G(V, V') \cong \oplus_{t=1}^s \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}^{m(\omega_t,V)}, \mathbb{C}^{m(\omega_t,V')}) \tag{2.4}$$

Going back to our fixed vector space W, we already know that the space $\otimes^l W$, $l > 0$ is a G-representation as well and, as such,

$$\otimes^l W \cong \oplus_{t=1}^s N_{\omega_t} \otimes \mathbb{C}^{m(\omega_t,\otimes^l W)}.$$

We will denote by $\mathbf{m}(l)$ the s-tuple

$$\mathbf{m}(l) = (m(\omega_1, \otimes^l W), \dots, m(\omega_s, \otimes^l W)).$$

In particular, $\mathbf{m}(1)$ will be denoted by $\mathbf{m} = (m_1, \dots, m_s)$. Moreover, if χ denotes the associated character to the representation $\rho : G \longrightarrow GL(W)$, the decomposition (2.3) above induces an equality of characters

$$\chi = \sum_{t=1}^{s} m_t \omega_t \quad m_t \in \mathbb{Z}.$$

If $\mathbf{a} = (a_t)_{t=1,\dots,s}, \mathbf{b} = (b_t)_{t=1,\dots,s} \in \mathbb{N}^s$, we write $\mathbf{a} \leq \mathbf{b}$ if $a_t \leq b_t$ for each $t = 1, \dots, s$. Similarly, $\min\{\mathbf{a}, \mathbf{b}\}$ is the s-tuple given by the minimum of each entry.

Lemma 2.11. *With this notation, we have* $\mathbf{m}(l) \leq \mathbf{m}(l')$ *if* $l \leq l'$.

PROOF. We prove that $\mathbf{m}(l) \leq \mathbf{m}(l+1)$ for any l. First of all, we show that if $\omega_1 \in \Omega$ is the trivial character, then $m_1 \geq 1$. To this aim, notice that the vector $\sum_{b \in B} b \in W$ is invariant by the action of any $g \in G$. In particular, we have $\sum_{b \in B} b \in W[\omega_1]$ and so ω_1 does appear in the decomposition of χ with non-zero coefficient. Now, given $l > 0$, write $\chi^l = \sum_t a_t \omega_t$. The claim follows from the fact that the coefficient of any irreducible character of G, say ω_t, in χ^{l+1} is just $m_1 a_t + \ldots \geq a_t$. □

Notation 2.12. Following [3], if $\mathbf{m}$, $\mathbf{n}$ are s-tuples of positive integers we will use the notation $M_{\mathbf{m},\mathbf{n}}$ to denote the space $M_{m_1,n_1} \times \cdots \times M_{m_s,n_s}$ and if $A = (A_1, \dots, A_s) \in M_{\mathbf{m},\mathbf{n}}$, we will write

$$\mathbf{rk}\,(A) = (\mathrm{rk}\,(A_1), \dots, \mathrm{rk}\,(A_s))\,.$$

Notice that $M_{\mathbf{m},\mathbf{n}}$ can be understood as the subspace of $M_{\sum m_t, \sum n_t}$ given by the block-diagonal matrices with blocks of sizes $m_t \times n_t$.

2.2. Flattenings and thin flattenings

The following definitions will be crucial for our purposes.

Definition 2.13. Let T be a phylogenetic tree on (G, W) and let $L_1 \mid L_2$ be a bipartition of its leaves. Let ψ be a G-tensor on T.

The *flattening of ψ along* $L_1 \mid L_2$, denoted by $flat_{L_1|L_2}\psi$, is the image of ψ via the isomorphism

$$\mathcal{L}(T)^G \cong \mathrm{Hom}_G\left(\otimes_{L_1} W, \otimes_{L_2} W\right).$$

The *thin flattening of ψ along* $L_1 \mid L_2$ is the s-tuple of linear maps, denoted by $Tf_{L_1|L_2}(\psi)$, obtained from $flat_{L_1|L_2}\psi$ via the isomorphism

$$\mathrm{Hom}_G(\otimes_{L_1} W, \otimes_{L_2} W) \cong \bigoplus_{t=1}^{s} \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}^{\mathbf{m}(l_1)_t}, \mathbb{C}^{\mathbf{m}(l_2)_t}).$$

Remark 2.14. Notice that if $\psi \in \mathcal{L}(T)^G$ and $L_1 \mid L_2$ is a bipartition of $L(T)$, then

$$\left(flat_{L_1|L_2}\psi\right)(u) = \langle \psi \mid u \rangle, \qquad \forall u \in \otimes_{L_1} W$$

where $\langle \cdot \mid \cdot \rangle$ is the operation defined in (2.2).

Notation 2.15. If $Tf_{L_1|L_2}(\psi) = (\psi_1, \psi_2, \ldots, \psi_s)$, we write

$$\mathbf{rk}\ Tf_{L_1|L_2}(\psi) = (\mathrm{rk}\,(\psi_1), \ldots, \mathrm{rk}\,(\psi_t)).$$

Notation 2.16. Given a phylogenetic tree T and an edge $e \in E(T)$, we denote $\mathbf{1}_e = \sum_{i,j=1}^{k} b_i \otimes b_j \in W_{e_0} \otimes W_{e_1}$. Similarly, we denote $\mathrm{id}_e = \sum_{i=1}^{k} b_i \otimes b_i \in W_{e_0} \otimes W_{e_1}$. We write $\mathbf{id}_T = (\mathrm{id}_e)_{e \in E(T)}$ and call it the *no-mutation presentation of* T.

3. The ideal of an equivariant model

In this section, we essentially prove that edge invariants are indeed phylogenetic invariants (see Introduction). The proof of this result is quite technical as it is valid for any equivariant model.

Let T be a trivalent phylogenetic tree T on (G, W). A *subtree* of T is a tree T' such that $L(T') \subset L(T)$ and $E(T') \subset E(T)$. Given a bipartition β of $L(T)$, it is clear that β induces a (possibly trivial) bipartition on the leaves of any subtree.

Write n_β for the maximal number m of disjoint subtrees $T_1, \ldots, T_m$ of T such that for every i, $\beta \cap L(T_i)$ is an edge split of T_i and $L(T) = \cup_{i=1}^{m} L(T_i)$. Notice that

$n_\beta = 1$ if and only if β is an edge split of T and otherwise, as long as $n \geq 4$, we will have $n_\beta \geq 2$. From now on, we will denote

$$\mathbf{m}_{\beta,T} = \mathbf{m}(n_\beta).$$

The main goal of this section is to prove the following Proposition, whose interest lies in the fact that it translates the topology of a tree into rank conditions of suitable matrices.

Proposition 3.1. *Let T be a trivalent phylogenetic tree T on (G, W) and let $\beta = L_1 \mid L_2$ be a bipartition of $L(T)$ as above. Then, we have*

$$\mathbf{rk}\ Tf_\beta(\psi) \leq \mathbf{m}_{\beta,T} \qquad \forall \psi \in V(T),$$

and there exists a non-empty Zariski open set $U_\beta \subset V(T)$ such that the equality holds for every $\psi \in U_\beta$. Moreover,

(i) β is an edge split in T if and only if $\mathbf{m}_{\beta,T} = \mathbf{m}$.

(ii) If β is not an edge split in T, then $\mathbf{m}_{\beta,T} \geq \mathbf{m}(2)$.

The existence of the Zariski open subset above where the flattening attains the expected rank cannot be proven by a simple dimension counting as the following example shows.

Example 3.2. Consider $G = \{\mathrm{id}\} \subset \mathfrak{S}_4$ and the quartet tree T having an inner edge e. Then $Tf_e(\psi)$ can be seen as a 16×16 matrix M and its expected rank is 4 according to Proposition 3.1(i). The variety $V_G(T)$ has dimension 60 and is contained in the determinantal variety defined by the 5×5 minors of M, which has dimension $256 - (16 - 5 + 1)(16 - 5 + 1) = 112$. A priori $V_G(T)$ could also be included in the variety of 4×4 minors of M which has dimension $256 - (16-4+1)(16-4+1) = 87$, so that a general element of $V_G(T)$ would not have the expected rank 4.

Remark 3.3. Notice that for the case of the general Markov model, this result provides a bound for the generic rank of the flattening along a bipartition which does not coincide with the bound provided in [13, Theorem 19.5]. As an example, consider the second tree in the figure 3 and the bipartition $\beta =$ “black”/“white” at its leaves. According to [13], the generic rank of the flattening along β of a tensor under the general Markov model on this tree should be k^4, while according to our Proposition 3.1 it is k^3.

Before proving Proposition 3.1, we state a couple of lemmas.

Lemma 3.4. *Let T be a trivalent phylogenetic tree and let $\beta = L_1 \mid L_2$ be an edge split of T. For a generic evolutionary presentation A of T, it holds that* $\operatorname{rk}\, flat_\beta(\Psi_T(A)) = k$.

Proof. Let $e \in E(T)$ be the edge corresponding to the bipartition β. Insert a vertex p in e and decompose $T = T_1 *_p T_2$, T_i with leaves $L_i \cup \{p\}$. Define $\psi_1 = \Psi_{T_1}(A_1)$ $\psi_2 = \Psi_{T_2}(A_2)$, where $A_1 \in \operatorname{Par}(T_1)$ and $A_2 \in \operatorname{Par}(T_2)$ are the evolutionary presentations obtained from A as: $(A_i)_{e'} = A_{e'}$ if e' is an edge different than e, $i = 1, 2$, $(A_1)_{e_p^1} = id$ if e_p^1 is the edge of T_1 containing p and $(A_2)_{e_p^2} = A_e$ if e_p^2 is the edge of T_2 containing p. Then,

$$\Psi_T(A) = \psi_1 *_p \psi_2 = \sum_{j=1}^{k} \langle \psi_1 \mid b_j \rangle \otimes \langle \psi_2 \mid b_j \rangle.$$

If $\mathbf{x}_1 \in \otimes_{L_1} W$, we derive that

$$\langle \Psi_T(A) \mid \mathbf{x}_1 \rangle = \sum_{j=1}^{k} (\psi_1 \mid \mathbf{x}_1 \otimes b_j)\, \langle \psi_2 \mid b_j \rangle,$$

which can be written as the composition of the maps

$$\begin{array}{rcl} flat_{L_1|\{p\}}(\psi_1) : \otimes_{L_1} W & \rightarrow & W_p \\ \mathbf{x}_1 & \mapsto & \displaystyle\sum_j (\psi_1 \mid \mathbf{x}_1 \otimes b_j)\, b_j \end{array}$$

and

$$\begin{array}{rcl} flat_{\{p\}|L_2}(\psi_2) : W_p & \rightarrow & \otimes_{L_2} W \\ x & \mapsto & \langle \psi_2 \mid x \rangle \end{array}$$

From this, it follows that $flat_\beta(\Psi_T(A))$ factorizes through W_p and, in particular, $\operatorname{rk}\, flat_\beta(\Psi_T(A)) \leq k$. Since the map $\Psi_T : \operatorname{Par}(T) \rightarrow \mathcal{L}(T)$ is continous, the condition $\operatorname{rk}\, flat_\beta \Psi_T(A) \geq k$ defines an open set U_β in $\operatorname{Par}(T)$. To finish the proof, we only need to prove that the open set U_β is not empty. To this aim, it is enough to consider the no-mutation presentation $\mathbf{id}_T = (id_e)_{e \in E(T)}$. Clearly, the linear map

$$flat_\beta \Psi_T(\mathbf{id}_T) : b_{i_1} \otimes \ldots \otimes b_{i_{l_1}} \quad \mapsto \quad \begin{cases} 0 & \text{if } b_{i_j} \neq b_{i_k} \\ \overbrace{b_i \otimes \ldots \otimes b_i}^{l_2} & \text{if } b_{i_j} = b_i, \forall j \end{cases}$$

has rank equal to k, so $\mathbf{id}_T \in U_\beta$. $\square$

Lemma 3.5. *Let T be a trivalent phylogenetic tree and let $q \in L(T)$. Let T' be the phylogenetic tree obtained from T by removing q and the pendant edge $e \in E(T)$ adjacent to it. Let $A \in \mathrm{Par}(T)$ be such that $A_e = \mathbf{1}_e$ (see notation 2.16). Then, for any $b_i \in B$ (the basis at W_q), it holds*

$$\langle \Psi_T(A) \mid b_i \rangle = \Psi_{T'}(A'),$$

where $A' \in \mathrm{Par}(T')$ is the restriction of A to T'.

Proof. Let $p \in V(T)$ be the adjacent vertex to q, so that $T = T_e *_p T_1 *_p T_2$ and T_e is the 1-edge tree with vertices q and p. Then, $\Psi_T(A)$ decomposes as

$$\begin{aligned} \Psi_T(A) &= \Psi_{T_e}(\mathbf{1}_e) *_p \Psi_{T_1}(A) *_p \Psi_{T_2}(A) = \\ &= \sum_j \langle \mathbf{1}_e \mid b_j \rangle \otimes \langle \Psi_{T_1}(A) \mid b_j \rangle \otimes \langle \Psi_{T_2}(A) \mid b_i \rangle. \end{aligned}$$

Since $\langle \mathbf{1}_e \mid b_j \rangle = \sum_t b_t$, we derive that

$$\begin{aligned} \langle \Psi_T(A) \mid b_i \rangle &= \sum_j (\sum_t b_t \mid b_i) \langle \Psi_{T_1}(A) \mid b_j \rangle \otimes \langle \Psi_{T_2}(A) \mid b_i \rangle = \\ &= \sum_j \langle \Psi_{T_1}(A) \mid b_j \rangle \otimes \langle \Psi_{T_2}(A) \mid b_i \rangle = \\ &= \Psi_{T_1}(A) *_p \Psi_{T_2}(A) = \Psi_{T'}(A') \end{aligned}$$

and the claim follows. $\square$

Proof of Proposition 3.1. We proceed in 3 steps.

Step 1. We show that for any evolutionary presentation $A \in \mathrm{Par}_G(T)$, we have

$$\mathrm{rk}\ flat_\beta(\Psi_T(A)) \leq k^{n_\beta}.$$

To prove this bound, we decompose the tree T in the following way (see figure 2): write $T_{(j)}$, $j = 1, \ldots, n_\beta$ for a maximal collection of subtrees of T such that $L(T) = \cup_{j=1}^{n_\beta} L(T_{(j)})$ and the bipartitions $\beta_{(j)}$ induced by β at the leaves of $T_{(j)}$ are edge splits of $T_{(j)}$. We assume that we have ordered the subtrees $T_{(j)}$ such that $T \setminus T_{(1)}$ is connected and $T_{(2)}$ is joined to $T_{(1)}$ by an edge. For every j, write $e_{(j)}$

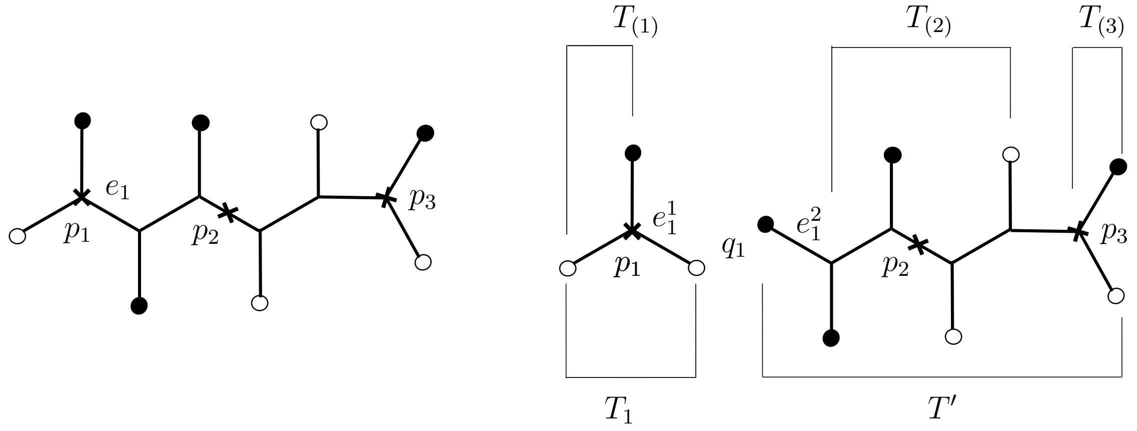


Figure 2: Decomposition of the tree T on the left in the proof of Proposition 3.1.

for the edge of $T_{(j)}$ giving this split and insert a new vertex, say p_j, in it. Then, we will show that $flat_\beta(\Psi_T(A))$ factorizes through $\otimes_{j=1}^{n_\beta} W_{p_j}$. To this aim, we will use induction on n_β.

If $n_\beta = 1$, the claim follows by the proof of Lemma 3.4 and there is nothing else to prove. For the general case, let e_1 be the edge of T adjacent to $T_{(1)}$. By inserting a vertex q_1 in e_1, we decompose e_1 into two edges e_1^1 and e_1^2 and write $T = T_1 *_{q_1} T'$, where T_1 is obtained by adding e_1^1 as a pendant edge to $T_{(1)}$ and T' is the remaining tree. We assume that q_1 belongs to the same connected component as $L_1 \cap T_{(2)}$ when removing $e_{(2)}$ from T (we would proceed similarly if q_1 belonged to the component of $L_2 \cap T_{(2)}$) and we write $\beta' = L_1' \mid L_2'$ where $L_1' = (L_1 \cap L(T')) \cup \{q_1\}$ and $L_2' = L_2 \cap L(T')$.

Define $\psi_1 = \Psi_{T_1}(A_1)$ where $(A_1)_e = A_e$, if $e \neq e_1^1$ and $(A_1)_{e_1^1} = \mathrm{id}_{e_1^1}$. Similarly, define $\psi' = \Psi_{T'}(A')$ where $(A')_e = A_e$ if $e \neq e_1^2$ and $(A')_{e_1^2} = A_{e_1}$. Then, the decomposition of T above induces a decomposition of $\psi = \Psi_T(A)$ as

$$\psi = \psi_1 *_{q_1} \psi' = \sum_{j=1}^{k} \langle \psi_1 \mid b_j \rangle \otimes \langle \psi' \mid b_j \rangle. \tag{3.1}$$

Decompose $T_1 = T_{(1)}^1 *_p T_{(1)}^2 *_p T_{e_1^1}$, where p is the node in $T_{(1)}$ adjacent to q_1, $T_{(1)}^1$ is the tree determined by the leaves of $L_1^1 := L_1 \cap L(T_1)$ and similarly, $T_{(1)}^2$ is the tree determined by the leaves of $L_2^1 := L_2 \cap L(T_1)$. Notice that $T_{(1)}^1 * T_{(1)}^2 = T_{(1)}$. Write $\psi_{(1)}^1 \in \mathcal{L}(T_{(1)}^1)$ and $\psi_{(1)}^2 \in \mathcal{L}(T_{(1)}^2)$ for the images by $\Psi_{T_{(1)}^1}$ and $\Psi_{T_{(1)}^2}$ of the corresponding restrictions of $A_1 \in \mathrm{Par}(T_{(1)})$. Then,

$$\psi_1 = \psi^1_{(1)} * \psi^2_{(1)} * \mathrm{id}_{e_1} = \sum_{i=1}^{k} \langle \psi^1_{(1)} \mid b_i \rangle \otimes \langle \psi^2_{(1)} \mid b_i \rangle \otimes b_i, \tag{3.2}$$

so $\langle \psi_1 \mid b_j \rangle = \langle \psi^1_{(1)} \mid b_j \rangle \otimes \langle \psi^2_{(1)} \mid b_j \rangle$. Thus, from (3.1) we have

$$\psi = \sum_{j=1}^{k} \langle \psi^1_{(1)} \mid b_j \rangle \otimes \langle \psi^2_{(1)} \mid b_j \rangle \otimes \langle \psi' \mid b_j \rangle.$$

Given $\mathbf{x}_1 \in \otimes_{L_1^1} W$ and $\mathbf{x}' \in \otimes_{L_1 \cap T'} W$, we have

$$\begin{aligned} flat_\beta(\psi)(\mathbf{x}_1 \otimes \mathbf{x}') &= \sum_{j=1}^{k} \langle \psi^1_{(1)} \mid \mathbf{x}_1 \otimes b_j \rangle \otimes \langle \psi^2_{(1)} \mid b_j \rangle \otimes \langle \psi' \mid b_j \otimes \mathbf{x}' \rangle = \\ &= \sum_{j=1}^{k} \big((\psi^1_{(1)} \mid \mathbf{x}_1 \otimes b_j) \langle \psi^2_{(1)} \mid b_j \rangle \big) \otimes \langle \psi' \mid b_j \otimes \mathbf{x}' \rangle \qquad (3.3) \end{aligned}$$

where the last equality holds because $\langle \psi^1_{(1)} \mid \mathbf{x}_1 \otimes b_j \rangle \in \mathbb{C}$.

By the induction hypothesis, we know that the map

$$flat_{\beta'}(\psi') : b_j \otimes \mathbf{x}' \mapsto \langle \psi' \mid b_j \otimes \mathbf{x}' \rangle$$

factorizes through $\otimes_{t=2}^{n_\beta} W_{p_j}$, that is, there exist homomorphisms

$$H'_1 : \otimes_{L'_1} W \to \otimes_{t=2}^{n_\beta} W_{p_j} \qquad \text{and} \qquad H'_2 : \otimes_{t=2}^{n_\beta} W_{p_j} \to \otimes_{L'_2} W$$

such that $flat_{\beta'}(\psi') = H'_2 \circ H'_1$. To show that $flat_\beta(\psi)$ factorizes through $\otimes_{t=1}^{n_\beta} W_{p_j}$, consider the map

$$\begin{aligned} H_1 : \otimes_{L_1} W &\to W_{p_1} \otimes (\otimes_{t=2}^{n_\beta} W_{p_j}) \\ \mathbf{x}_1 \otimes \mathbf{x}' &\mapsto \sum_{j=1}^{k} (\psi^1_{(1)} \mid \mathbf{x}_1 \otimes b_j) b_j \otimes H_1(b_j \otimes \mathbf{x}') \end{aligned}$$

and compose it with $H_2 = flat_{\{p_1\}|L_2^1}(\psi^2_{(1)}) \otimes H'_2$:

$$\begin{aligned} H_2 : W_{p_1} \otimes (\otimes_{t=2}^{n_\beta} W_{p_j}) &\to \otimes_{L_2} W \\ x_1 \otimes (x_2 \otimes \ldots \otimes x_{n_\beta}) &\mapsto \langle \psi^2_{(1)} \mid x_1 \rangle \otimes H'_2(x_2 \otimes \ldots x_{n_\beta}). \end{aligned}$$

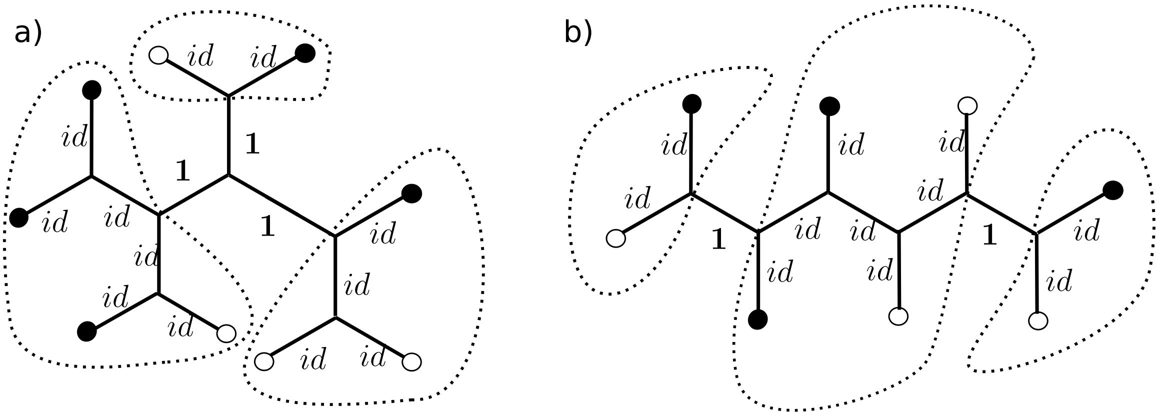


Figure 3: The bipartition β induces a split on the subtrees surrounded with a doted line. The number n_β is 3 in both trees. The evolutionary presentations provided here are those used in the proof of Proposition 3.1.

It is straightforward to check that this composition of maps applied to $\mathbf{x}_1 \otimes \mathbf{x}'$ equals the expression obtained in (3.3). This proves the claim of the *Step 1*.

Once we know that the rank of $flat_\beta \Psi_T(A)$ is upper bounded by k^{n_β}, the condition rk $flat_\beta \Psi_T(A) = k^{n_\beta}$ becomes equivalent to the condition rk $flat_\beta(\Psi_T(A)) \geq k^{n_\beta}$, which defines an open set U_β in $\mathrm{Par}(T)$.

Step 2. The next step is to show that the open set U_β is non-empty. To this aim, take $\varphi_T = \Psi_T(A_T^0)$ where $A_T^0 \in \mathrm{Par}_G(T)$ is given by

$$(A_T^0)_e = \begin{cases} \mathrm{id}_e & \text{if } e \in E(T_{(j)}) \text{ for some } j; \\ \mathbf{1}_e & \text{otherwise.} \end{cases}$$

The figure 3 shows two examples of this evolutionary presentation.

We use induction on n_β to show that rk $flat_\beta(\varphi_T) = k^{n_\beta}$. If $n_\beta = 1$, we are in the situation described in Lemma 3.4 and the claim follows from the proof given there. For the general case, keep the notation introduced in *Step 1.* The decomposition (3.1) is

$$\psi = \Psi_{T_1}(\mathbf{id}_{T_1}) *_{q_1} \psi' = \sum_i \langle \Psi_{T_1}(\mathbf{id}_{T_1}) \mid b_i \rangle \otimes \langle \psi' \mid b_i \rangle.$$

where $\psi' \in \mathcal{L}(T')$ is the image by $\Psi_{T'}$ of the restriction A' of A^0 to the edges of T'. Now, notice that $A'_{e_1^2} = \mathbf{1}_{e_1^2}$, so Lemma 3.5 applies to ψ' and we infer that for any

$b_i \in W_{q_1}$, it holds

$$\langle \psi' \mid b_i \rangle = \varphi_{T''},$$

where T'' is the tree obtained from T' by removing the edge e_1^2 and $\varphi_{T''}$ is the image by $\Psi_{T''}$ of an evolutionary presentation as the one described above, that is, $\varphi_{T''} = \Psi_{T''}(A^0_{T''})$. Putting all together, we obtain

$$\begin{aligned} flat_\beta(\psi)(\mathbf{x}_1 \otimes \mathbf{x}') &= \sum_{i=1}^{k} \langle \Psi_{T_1}(\mathbf{id}_{T_1}) \mid \mathbf{x}_1 \otimes b_i \rangle \otimes \langle \varphi_{T''} \mid \mathbf{x}' \rangle = \\ &= \left(\sum_{i=1}^{k} \langle \Psi_{T_1}(\mathbf{id}_{T_1}) \mid \mathbf{x}_1 \otimes b_i \rangle \right) \otimes \langle \varphi_{T''} \mid \mathbf{x}' \rangle \end{aligned}$$

On the one hand, if $\mathbf{x}_1 = b_j \otimes \ldots \otimes b_j$, it is clear that

$$\sum_i \langle \Psi_{T_1}(\mathbf{id}_{T_1}) \mid \mathbf{x}_1 \otimes b_i \rangle = b_j \otimes \ldots \otimes b_j,$$

and the left term in this equality is 0 if $\mathbf{x}_1$ is an element in the basis of $\otimes_{L_1^1} W$ different than $b_j \otimes \ldots \otimes b_j$. Therefore the rank of the map $\mathbf{x}_1 \mapsto \sum_i \langle \Psi_{T_1}(\mathbf{id}_{T_1}) \mid \mathbf{x}_1 \otimes b_i \rangle$ is k.

On the other hand, the induction hypothesis implies that rk $flat_{\beta''}(\varphi'') = k^{n_\beta - 1}$ where $\beta'' = L_1 \cap L(T'') \mid L_2 \cap L(T'')$. From this, we derive that the rank of $flat_\beta(\psi)$ equals $k \times k^{n_\beta - 1} = k^{n_\beta}$.

This proves the claim and so, the open set $U_\beta \subset \mathrm{Par}(T)$ defined above is non-empty.

Step 3. To finish the proof, notice that the presentation $A^0 \in \mathrm{Par}(T)$ defined in *Step 2* is equivariant for the whole group $\mathfrak{S}_k$. Therefore, once a subgroup $G \subset \mathfrak{S}_k$ is given, the restriction of U_β to $\mathrm{Par}_G(T)$ is non-empty. On the other hand, if $A \in \mathrm{Par}_G(T)$, then $\psi = \Psi_T(A) \in \mathcal{L}(T)^G$, $flat_\beta(\psi)$ is G-equivariant and all the maps are G-equivariant homomorphisms, so if $\psi = \Psi_T(A)$ for some $A \in \mathrm{Par}_G(T)$, then $\mathrm{Im} flat_\beta(\psi)$ is a G-representation isomorphic to some quotient of $\otimes^{n_\beta} W$. From this, we infer that the decomposition of $\mathrm{Im} flat_\beta(\psi)$ into isotypic components

$$\mathrm{Im} flat_\beta(\psi) \cong \oplus_{i=1}^{s} N_{\omega_i} \otimes \mathbb{C}^{d_i}$$

satisfies that $d_i \leq m_i(n_\beta)$, for $i = 1, \ldots, s$. Therefore,

$$\mathbf{rk}\ Tf_\beta(\psi) = (d_1, \ldots, d_s) \leq \mathbf{m}(n_\beta)$$

and the equality holds if and only if rk $flat_\beta(\psi) = k^{n_\beta}$.

To conclude, if β is an edge split, then $n_\beta = 1$ $\mathbf{m}_{\beta,T} = \mathbf{m}$. This proves (i). If β is not an edge split, it is clear that $n_\beta \geq 2$ and the claim of (ii) follows by Lemma 2.11. □

Remark 3.6. The preceeding proof actually shows that the dense open set $U_\beta \subset \mathrm{Par}_G(T)$ cuts the set of stochastic parameters, i.e

$$U_\beta \cap \prod_{e \in E(T)} \Delta^G \neq \emptyset,$$

where Δ^G is the set of Markov matrices, that is, matrices whose entries are all non-negative and whose columns sum to 1. Keeping the notations introduced in the proof, it is enough to take $A \in \mathrm{Par}_G(T)$ given by

$$A_e = \begin{cases} \mathrm{id}_e & \text{if } e \in E(T_{(j)}) \text{ for some } j; \\ \frac{1}{4}(\mathbf{1}_e) & \text{otherwise.} \end{cases}$$

Proposition 3.1 suggests the following definitions.

Definition 3.7. If $L_1 \mid L_2$ is a bipartition of $L(T)$, the *ideal of* $L_1 \mid L_2$, denoted by $I_{L_1|L_2}$, is the ideal in the coordinate ring of $\mathcal{L}(T)^G$ defined by the conditions

$$\mathbf{rk}\ Tf_{L_1|L_2}(\psi) \leq \mathbf{m},$$

$\psi \in \mathcal{L}(T)^G$ being a tensor of indeterminates. Equivalently, $I_{L_1|L_2}$ is generated by the (m_t+1)-minors of the t-th box of $Tf_{L_1|L_2}(\psi) \in M_{\mathbf{m}(l_1),\mathbf{m}(l_2)}$, for $t = 1, \ldots, s$ (see Notation 2.15).

Notation 3.8. Let T be a phylogenetic tree on (G, W) and let e be an edge of T that splits the leaves into two sets L_1 and L_2 of cardinality l_1 and l_2, respectively. The ideal $I_{L_1|L_2}$ will be also denoted as I_e. Due to Proposition 3.1 we have that if e belongs to $E(T)$, then $I_e \subseteq \mathcal{I}(T)$.

Definition 3.9. The *edge invariants* of T are the elements of the ideal $\sum_{e \in E(T)} I_e$.

Proposition 3.1 proves that edge invariants are *phylogenetic invariants*, that is, elements in $\mathcal{I}(T)$ that do not vanish on all points of $\cup_T V(T)$ where the union runs over all trivalent tree topologies. Indeed, given a phylogenetic tree T_0 on (G, W)

and an edge $e \in E(T_0)$, there exist trivalent trees that do not have e as an edge split and so I_e is not contained in $\mathcal{I}(\cup_T V(T))$.

Is is worth highlighting that using Proposition 3.1 we also obtain the *generic identifiability* of the tree topology for equivariant models. The tree topology of a model of sequence mutation is said to be generically identifiable if for generic choices of stochastic parameters $A \in \prod_{e \in E(T)} \Delta^G, A' \in \prod_{e \in E(T')} \Delta^G$ (see Remark 3.6), $\Psi_T(A) = \Psi_{T'}(A')$ implies $T = T'$ (see for instance [2]). In order to prove this kind of results, one only has to show the corresponding irreducible varieties $V(T)$ and $V(T')$ are not contained one into the other. We obtain the following result that was already known for the general Markov model (see [23]) and for group-based models [24].

Corollary 3.10. *The tree topology is generically identifiable in all equivariant evolutionary models.*

Proof. Let T, T' be two different trivalent phylogenetic trees on (G, W). Then there is an edge split e in T that is not an edge split in T'. By Proposition 3.1, there exists an element f in I_e (and therefore in $\mathcal{I}(T)$) that does not belong to $\mathcal{I}(T')$. In terms of varieties this proves that $V(T') \subsetneq V(T)$, and $V(T) \subsetneq V(T')$ is proven similarly. As $V(T)$ and $V(T')$ are irreducible varieties, this shows that they meet properly. □

4. Phylogenetic Invariants

The purpose of this section is to prove that, for phylogenetic reconstruction, the only relevant invariants are the edge invariants introduced in the previous section. This is a natural result if one takes into account the Splits Equivalence Theorem in combinatorics (see Theorem 4.1 below). Let $\mathcal{T}$ be the set of isomorphism classes of trivalent tree topologies with leaf set $L = \{v_1, v_2, \ldots, v_n\}$. Two bipartitions $L_1|L_2$, $M_1|M_2$ of a set L are said to be *compatible* if at least one of the four intersections $L_1 \cap M_1$, $L_1 \cap M_2$, $L_2 \cap M_1$, $L_2 \cap M_2$ is empty. For example, if $L_1|L_2$, $M_1|M_2$ are two edge splits of the same tree T, then they are compatible. We recall that any trivalent tree on n leaves has $2n - 3$ interior edges.

Theorem 4.1 ([6], [21, Theorem 2.35]). *A collection $\mathcal{B}$ of $2n-3$ bipartitions of L is pairwise compatible if and only if there exists a tree $T \in \mathcal{T}$ such that $\mathcal{B}$ is the set of edge splits of T. Moreover, if such a tree T exists then it is unique.*

In order to make our result concerning phylogenetic invariants more precise we need to introduce some notation.

We fix $G \subset \mathfrak{S}_k$ and W as in section 2 and each topology $T \in \mathcal{T}$ will be considered as a phylogenetic tree on (G, W). Then all trees T in $\mathcal{T}$ have the same space of G-tensors which will be denoted by $\mathcal{L} = (\bigotimes_{i=1}^{n} W)^G$.

Definition 4.2. Let $\mathbf{o}$ be an s-tuple and let $\beta = L_1 \mid L_2$ be a bipartition of $\{v_1, v_2, \dots, v_n\}$. Then we let $D^{\beta}_{\leq \mathbf{o}}$ be the subvariety of $\mathcal{L}$ defined as

$$D^{\beta}_{\leq \mathbf{o}} = \{\psi \in \mathcal{L} \mid \mathbf{rk}\ Tf_{\beta}(\psi) \leq \mathbf{o}\}$$

and, if the thin flattening of $\psi \in \mathcal{L}$ is $Tf_{\beta}(\psi) = (\psi_1, \psi_2, \dots, \psi_s)$, we define $D^{\beta}_{<\mathbf{o}}$ to be the set

$$D^{\beta}_{<\mathbf{o}} = \{\psi \in \mathcal{L} \mid \text{rk } \psi_j < o_j \text{ for some j }\}.$$

For example, $D^{\beta}_{\leq \mathbf{m}}$ coincides with the set of zeroes $Z(I_{L_1,L_2})$. Notice that both $D^{\beta}_{\leq \mathbf{o}}$ and $D^{\beta}_{<\mathbf{o}}$ are algebraic sets although the second is not always irreducible.

Notation 4.3. Given a tree $T \in \mathcal{T}$ and using the notation of Proposition 3.1, for each bipartition $\beta = L_1 \mid L_2$ of $\{v_1, v_2, \dots, v_n\}$, we call $\mathbf{m}_{\beta,T}$ the maximum rank that $Tf_{\beta}(\psi)$ can have if ψ belongs to $V(T)$. Then Proposition 3.1 shows that

$$V(T) \subseteq D^{\beta}_{\leq \mathbf{m}_{\beta,T}}$$

and that $V(T) \setminus D^{\beta}_{<\mathbf{m}_{\beta,T}}$ is a dense open subset of $V(T)$ for any bipartition $\beta = L_1 \mid L_2$. We call this open subset $U_{T,\beta}$, so that $U_{T,\beta} = V(T) \setminus D^{\beta}_{<\mathbf{m}_{\beta,T}}$ is the locus of tensors $\psi \in V(T)$ that satisfy $\mathbf{rk}\ Tf_{\beta}(\psi) = \mathbf{m}_{\beta,T}$. We define $U_T = \cap_{\beta} U_{T,\beta}$, where the intersection is taken among all bipartitions of $\{v_1, v_2, \dots, v_n\}$. As $V(T)$ is an irreducible variety, U_T is still a dense open subset of $V(T)$ and it corresponds to the set of points in $V(T)$ whose flattening $Tf_{\beta}(\psi)$ along any partition β of the set of leaves of T has the expected rank $\mathbf{m}_{\beta,T}$.

With this set up in mind, the main result of this paper is the following.

Theorem 4.4. *For each $T \in \mathcal{T}$, let $U_T \subset V(T)$ be the dense open set defined above. Let p be a point in $\bigcup_{T \in \mathcal{T}} U_T \subseteq \mathcal{L}$ and let T_0 be any tree in $\mathcal{T}$. Then, p belongs to $V(T_0)$ if and only if p belongs to the set of zeroes $Z(\sum_{e \in E(T_0)} I_e)$.*

Remark 4.5. As we pointed out in the introduction, this result says that for a general point on $\bigcup_{T\in\mathcal{T}} V(T)$, it is enough to evaluate the edge invariants to decide to which variety $V(T)$ the point actually belongs to.

This result would still hold for non-trivalent trees when imposing that all trees in the corresponding set $\mathcal{T}$ have the same collection of degrees at interior vertices.

After all the technical issues in section 3, the proof of Theorem 4.4 is now straightforward.

PROOF OF 4.4. By Proposition 3.1 we already know that $\sum_{e\in E(T_0)} I_e \subseteq \mathcal{I}(T_0)$, therefore if $p \in V(T_0)$, we immediately have that p belongs to $Z(\sum_{e\in E(T_0)} I_e)$.

Conversely, let $p \in \cup_{T\in\mathcal{T}} U_T$. Then p belongs to $U_T \subset V(T)$ for a certain $T \in \mathcal{T}$, so that $\mathbf{rk}\ Tf_\beta(p) = \mathbf{m}_{\beta,T}$ for any bipartition β of $\{v_1, v_2, \ldots, v_n\}$. On the other hand, if $p \in Z(\sum_{e\in E(T_0)} I_e)$, then $p \in Z(I_e)$ for any $e \in E(T_0)$ and hence, $\mathbf{rk}\ Tf_e(p) \leq \mathbf{m}$ for all $e \in E(T_0)$. This implies that $\mathbf{m}_{e,T} \leq \mathbf{m}$ for all $e \in E(T_0)$, which can only happen if e is a split of T for all $e \in E(T_0)$ (see Proposition 3.1). But two trivalent trees T and T_0 on n leaves have the same collection of splits if and only if $T = T_0$ (see Theorem 4.1), so the proof is concluded. □

Remark 4.6. The proof of Theorem 4.4 also shows that the intersection $U_T \cap U_{T'}$ is empty for any $T \neq T' \in \mathcal{T}$. However, there exists points in $V(T) \cap V(T')$ for any $T \neq T'$. Indeed, it is enough to consider $\psi_T(A)$ where A is the no-mutation presentation; then $\psi_T(A)$ lies in $V(T')$ for all T'. This proves that $\bigcap_T V(T)$ is not empty but one can also prove that, if $n \geq 5$, for any two different tree topologies T_1, T_2 one has $V(T_1) \cap V(T_2) \neq \bigcap_T V(T)$.

In the next Corollary we give an open subset $\mathcal{U}$ defined intrinsically from the ambient space $\mathcal{L}$ such that $\mathcal{U} \cap \cup_T V(T) = \cup_T U_T$. This is relevant for biological applications because then we will be able to check whether the given data point lies (or rather *is close to*) in $\cup_T U_T$. From now on let $\mathcal{B}$ be the set of all bipartitions of $\{v_1, \ldots, v_n\}$.

Corollary 4.7. *Let* $\mathcal{U} = \bigcup\limits_{T\in\mathcal{T}} \bigcap\limits_{\beta\in\mathcal{B}} (\mathcal{L} \setminus D^{\beta}_{<\mathbf{m}_{\beta,T}})$. *Then*

$$\mathcal{U} \cap \bigcup_{T\in\mathcal{T}} V(T) = \bigcup_{T\in\mathcal{T}} U_T$$

and if p is a point in $\mathcal{U} \cap \bigcup_{T\in\mathcal{T}} V(T)$ and T_0 is any tree in $\mathcal{T}$, then p belongs to $V(T_0)$ if and only if p belongs to the set of zeroes $Z(\sum_{e\in E(T_0)} I_e)$.

Proof. We just need to prove that $\mathcal{U} \cap (\bigcup_{T\in\mathcal{T}} V(T)) = \bigcup_{T\in\mathcal{T}} U_T$ because the other assertion follows from Theorem 4.4.

We have $\mathcal{U} \cap (\bigcup_{T\in\mathcal{T}} V(T)) = \bigcup_{T,T'} V(T) \cap (\cap_\beta \mathcal{L} \setminus D^{\beta}_{<\mathbf{m}_{\beta,T'}})$. If $T \neq T'$ this intersection is the empty set as we can see taking β an edge split of T but not of T'. Hence we obtain $\mathcal{U} \cap (\cup_{T\in\mathcal{T}} V(T)) = \bigcup_T V(T) \cap (\bigcap_\beta \mathcal{L} \setminus D^{\beta}_{<\mathbf{m}_{\beta,T}})$, which is precisely $\cup_T U_T$. □

In terms of ideals, Theorem 4.4 says the following:

Corollary 4.8. *Let R be the polynomial ring of $\mathcal{L}$ and let f be any element in*

$$\left(\sum_{T\in\mathcal{T}} \bigcap_{\beta\in\mathcal{B}} \mathcal{I}(D_{<\mathbf{m}_{\beta,T}})\right) \setminus \bigcap_T \mathcal{I}(T).$$

Then, the following equality holds in the localized ring $(R/\bigcap_T \mathcal{I}(T))_{\overline{f}}$

$$\left(\mathcal{I}(T_0)/\bigcap_T \mathcal{I}(T)\right)_{\overline{f}} = \left(rad\left(\sum_{e\in E(T_0)} I_e\right)/\bigcap_T \mathcal{I}(T)\right)_{\overline{f}}.$$

Proof. If we are given an f as above, then $U_f := \mathcal{L} \setminus \{f = 0\}$ is contained inside the open set $\mathcal{U}$ defined in Corollary 4.7. Indeed, an f as above is contained inside $rad(\sum_{T\in\mathcal{T}} \cap_\beta \mathcal{I}(D_{<\mathbf{m}_{\beta,T}}))$ which is equal to $\mathcal{I}(\cap_T \cup_\beta D^{\beta}_{<\mathbf{m}_{\beta,T}})$. Therefore $\cap_T \cup_\beta D^{\beta}_{<\mathbf{m}_{\beta,T}} \subset \{f = 0\}$ and $U_f \subset \mathcal{L} \setminus \cap_T \cup_\beta D^{\beta}_{<\mathbf{m}_{\beta,T}} = \mathcal{U}$.

In particular, $U_f \cap (\cup_T V(T))$ is contained inside $\cup_T U_T$. Therefore in U_f we still have that the variety $V(T_0)$ is defined inside $\cup_{T\in\mathcal{T}} V(T)$ by $\sum_{e\in E(T_0)} I_e$. Hence in terms of ideals in R_f we obtain the equality above. □

We do not know whether $\sum_{e\in E(T_0)} I_e$ is a radical ideal so we cannot remove *rad* from the expression above. We pose the following question:

Question 4.9. Given a set S of compatible splits, is $\sum_{\beta\in S} I_\beta$ radical?

Remark 4.10. In order to check whether Theorem 4.4 can be applied to a given data point $p \in \mathcal{L}$, it is enough to check that $f(p) \neq 0$ for a generic f in

$$\left(\sum_{T\in\mathcal{T}} \bigcap_{\beta\in\mathcal{B}} \mathcal{I}(D_{<\mathbf{m}_{\beta,T}})\right) \setminus \bigcap_T \mathcal{I}(T).$$

Such a polynomial f should be chosen a priori, so that when dealing with data one does not need to compute this ideal.

Remark 4.11. It is interesting to explore whether U_T can be defined by a complete intersection in the sense of [8]. This would reduce the number of generators of I_e to be used in phylogenetic reconstruction. However, this is another issue on which we plan to work in the future.

Although the degrees of a set of generators of the ideal of a phylogenetic tree evolving under the general Markov model or under the strand symmetric model are not known, Theorem 4.4 allows us to give the degrees of those invariants that are relevant in phylogenetic reconstruction. It is worth highlighting that these degrees do not depend on the number of leaves but only on the model and can be computed a priori (see the next sections for the precise examples of evolutionary models).

Corollary 4.12. *Let (G, W) be an equivariant evolutionary model and let $\mathbf{m} = (m_1, \ldots, m_s)$ be defined as in section 3. Then, for any tree topology on any number of leaves, the polynomials that are relevant for recovering the tree topology in phylogenetics have degrees in $\{m_1+1, \ldots, m_s+1\}$. In particular, the relevant phylogenetic invariants for the following evolutionary models have degrees:*

- *5 for the general Markov model.*
- *3 for the strand symmetric model.*
- *2 for the Kimura 3-parameter model.*
- *1 or 2 for the Kimura 2-parameter model.*
- *1 or 2 for the Jukes-Cantor model.*

5. Examples

In this section, we study some well-known evolutionary models in phylogenetics. Let $B = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ be the set of the four nucleotides and take $W = \langle \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T} \rangle_{\mathbb{C}} \cong \mathbb{C}^4$ with the bilinear form $(\cdot \mid \cdot)_W$ that makes B orthonormal. We consider the group of permutations of 4 elements,

$$\mathfrak{S}_4 = Sym\{B\}.$$

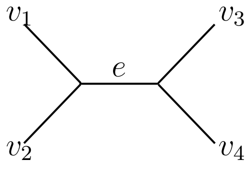
It is generated by $g_1 = (\mathrm{id}), g_2 = (\mathtt{AC}), g_3 = (\mathtt{ACG}), g_4 = (\mathtt{ACGT})$ and $g_5 = (\mathtt{AC})(\mathtt{GT})$, which correspond to the five conjugacy classes of $\mathfrak{S}_4$. We work with the natural permutation linear representation $\rho : \mathfrak{S}_4 \longrightarrow \mathrm{GL(W)}$ given by permuting the coordinates of W:

$$g_1 \mapsto \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad g_2 \mapsto \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad g_3 \mapsto \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$
$$g_4 \mapsto \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad g_5 \mapsto \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Write $\chi = \mathrm{Tr}(\rho(\cdot))$ for the character associated to it. We shall consider different subgroups of $\mathfrak{S}_4$, each one of them giving rise to a different equivariant model, according to the following diagram (we use the following shortenings: `GMM` for the general Markov model, `K81` for the Kimura 3 parameter model, `K80` for the Kimura 2 parameter model, `CS05` for the strand symmetric model and `JC69` for the Jukes-Cantor model):

$$\begin{tikzcd}
\{\mathrm{id}\} \arrow[d] & \texttt{GMM} \\
\langle(\texttt{AT})(\texttt{CG})\rangle \arrow[d] & \texttt{CS05} \arrow[u] \\
\langle(\texttt{AC})(\texttt{GT}), (\texttt{AG})(\texttt{CT})\rangle \arrow[d] & \texttt{K81} \arrow[u] \\
\langle(\texttt{ACGT}), (\texttt{AG})\rangle \arrow[d] & \texttt{K80} \arrow[u] \\
\mathfrak{S}_4 & \texttt{JC69} \arrow[u]
\end{tikzcd}$$

Our aim here is to describe in a unified fashion the edge invariants associated to these models for the case of a quartet tree topology T, with leaves v_1, v_2, v_3, v_4. Write $e = L_1 \mid L_2$ for the edge split corresponding to e, so that $L_1 = \{v_1, v_2\}$ and $L_2 = \{v_3, v_4\}$.



Remark 5.1. When the subgroup $G \subset \mathfrak{S}_4$ is abelian, the usual product of complex numbers induces on Ω_G a group structure. Then, if $\{u_1^t, \ldots, u_{m_t}^t\}$ is a basis for $W[\omega_t]$, for every $\omega_t \in \Omega_G$, we have that

$$\{u_{j_1}^{i_1} \otimes \cdots \otimes u_{j_l}^{i_l} \mid \omega_{i_1} \ldots \omega_{i_l} = \omega_t\}$$

is a $\mathbb{C}$-basis for $(\otimes^l W)[\omega_t]$.

5.1. General Markov model

As a first example, consider the trivial subgroup $\{\mathbf{id}\} \subset \mathfrak{S}_4$. The corresponding equivariant model is the *general Markov model*, which is the most general model in the Felsenstein hierarchy (see Ch.4 in [21]). Invariants for this model have been studied by Allman and Rhodes in [1, 3]. In this case, there is only one irreducible representation $\omega : G \to \mathbb{C}$ defined by mapping $(\mathbf{id})$ to 1. The character table is

$\Omega_{(1)}$	id
ω	1
χ	4

It follows that $\chi = 4\omega$. Keeping the notation introduced in 2.1, we have $\mathbf{m} = (4)$ and $W = W[\omega] \cong N_\omega \otimes \mathbb{C}^4$.

Now, for the case of four leaves, we have $\chi^2 = 16\omega$ and $\mathbf{m}(2) = (16)$. Then, the ideal I_e is defined by the condition

$$\mathbf{rk}\,(M) \leq (4)$$

where $M \in \mathrm{Hom}_G((W \otimes W)[\omega], (W \otimes W)[\omega]) \cong \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}^{16}, \mathbb{C}^{16})$ is a matrix of indeterminates whose columns and rows are indexed by the set $\{X_1 \otimes X_2\}_{X_1, X_2 \in B}$. The ideal I_e obtained by imposing the above rank condition is generated by $\binom{16}{5}\binom{16}{5}$ polynomials of degree 5.

5.2. Strand symmetric model

Take $G = \langle (\mathtt{AT})(\mathtt{CG}) \rangle$, which is isomorphic to $\mathbb{Z}/2\mathbb{Z}$. The equivariant matrices for this group have the following structure:

$$\begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}$$

The equivariant model associated to G is the *strand symmetric model* introduced in [10]. There are two irreducible characters ω_1, ω_2, and the character table is

Ω_G	id	$(\mathtt{AT})(\mathtt{CG})$
ω_1	1	1
ω_2	1	-1
χ	4	0

Notice that since G is abelian, all the irreducible representations have dimension one. It follows that $\chi = 2\omega_1 + 2\omega_2$. Thus, $\mathbf{m} = (2, 2)$ and we have a decomposition ([15, Corollary 2.14])

$$W = W[\omega_1] \oplus W[\omega_2],$$

where $W[\omega_1] \cong N_{\omega_1} \otimes \mathbb{C}^2$ and $W[\omega_2] \cong N_{\omega_2} \otimes \mathbb{C}^2$. Indeed, if we write

$$\mathtt{u}_1 = \mathtt{A} + \mathtt{T} \qquad \mathtt{u}_2 = \mathtt{C} + \mathtt{G} \qquad \mathtt{v}_1 = \mathtt{A} - \mathtt{T} \qquad \mathtt{v}_2 = \mathtt{C} - \mathtt{G},$$

we have

$$W[\omega_1] = \langle \mathtt{u}_1, \mathtt{u}_2 \rangle_{\mathbb{C}} \qquad W[\omega_2] = \langle \mathtt{v}_1, \mathtt{v}_2 \rangle_{\mathbb{C}}.$$

Now, we focus on the case of the tree with four leaves. We have $\chi^2 = 8\omega_1 + 8\omega_2$, so $\mathbf{m}(2) = (8, 8)$. Moreover, using that G is abelian (see Remark 5.1)

$$\begin{aligned}
W \otimes W[\omega_1] &= \langle \mathtt{u}_1 \otimes \mathtt{u}_1, \mathtt{u}_1 \otimes \mathtt{u}_2, \mathtt{u}_2 \otimes \mathtt{u}_1, \mathtt{u}_2 \otimes \mathtt{u}_2, \mathtt{v}_1 \otimes \mathtt{v}_1, \mathtt{v}_1 \otimes \mathtt{v}_2, \mathtt{v}_2 \otimes \mathtt{v}_1, \mathtt{v}_2 \otimes \mathtt{v}_2 \rangle \\
W \otimes W[\omega_2] &= \langle \mathtt{u}_1 \otimes \mathtt{v}_1, \mathtt{u}_1 \otimes \mathtt{v}_2, \mathtt{u}_2 \otimes \mathtt{v}_1, \mathtt{u}_2 \otimes \mathtt{v}_2, \mathtt{v}_1 \otimes \mathtt{u}_1, \mathtt{v}_1 \otimes \mathtt{u}_2, \mathtt{v}_2 \otimes \mathtt{u}_1, \mathtt{v}_2 \otimes \mathtt{u}_2 \rangle
\end{aligned}$$

Then, the ideal I_e is defined by the conditions

$$\mathbf{rk} \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \leq (2, 2)$$

where

$$M_1 = \begin{pmatrix}
q_{\mathtt{u_1u_1u_1u_1}} & q_{\mathtt{u_1u_1u_1u_2}} & q_{\mathtt{u_1u_1u_2u_1}} & q_{\mathtt{u_1u_1u_2u_2}} & q_{\mathtt{u_1u_1v_1v_1}} & q_{\mathtt{u_1u_1v_1v_2}} & q_{\mathtt{u_1u_1v_2v_1}} & q_{\mathtt{u_1u_1v_2v_2}} \\
q_{\mathtt{u_1u_2u_1u_1}} & q_{\mathtt{u_1u_2u_1u_2}} & q_{\mathtt{u_1u_2u_2u_1}} & q_{\mathtt{u_1u_2u_2u_2}} & q_{\mathtt{u_1u_2v_1v_1}} & q_{\mathtt{u_1u_2v_1v_2}} & q_{\mathtt{u_1u_2v_2v_1}} & q_{\mathtt{u_1u_2v_2v_2}} \\
q_{\mathtt{u_2u_1u_1u_1}} & q_{\mathtt{u_2u_1u_1u_2}} & q_{\mathtt{u_2u_1u_2u_1}} & q_{\mathtt{u_2u_1u_2u_2}} & q_{\mathtt{u_2u_1v_1v_1}} & q_{\mathtt{u_2u_1v_1v_2}} & q_{\mathtt{u_2u_1v_2v_1}} & q_{\mathtt{u_2u_1v_2v_2}} \\
q_{\mathtt{u_2u_2u_1u_1}} & q_{\mathtt{u_2u_2u_1u_2}} & q_{\mathtt{u_2u_2u_2u_1}} & q_{\mathtt{u_2u_2u_2u_2}} & q_{\mathtt{u_2u_2v_1v_1}} & q_{\mathtt{u_2u_2v_1v_2}} & q_{\mathtt{u_2u_2v_2v_1}} & q_{\mathtt{u_2u_2v_2v_2}} \\
q_{\mathtt{v_1v_1u_1u_1}} & q_{\mathtt{v_1v_1u_1u_2}} & q_{\mathtt{v_1v_1u_2u_1}} & q_{\mathtt{v_1v_1u_2u_2}} & q_{\mathtt{v_1v_1v_1v_1}} & q_{\mathtt{v_1v_1v_1v_2}} & q_{\mathtt{v_1v_1v_2v_1}} & q_{\mathtt{v_1v_1v_2v_2}} \\
q_{\mathtt{v_1v_2u_1u_1}} & q_{\mathtt{v_1v_2u_1u_2}} & q_{\mathtt{v_1v_2u_2u_1}} & q_{\mathtt{v_1v_2u_2u_2}} & q_{\mathtt{v_1v_2v_1v_1}} & q_{\mathtt{v_1v_2v_1v_2}} & q_{\mathtt{v_1v_2v_2v_1}} & q_{\mathtt{v_1v_2v_2v_2}} \\
q_{\mathtt{v_2v_1u_1u_1}} & q_{\mathtt{v_2v_1u_1u_2}} & q_{\mathtt{v_2v_1u_2u_1}} & q_{\mathtt{v_2v_1u_2u_2}} & q_{\mathtt{v_2v_1v_1v_1}} & q_{\mathtt{v_2v_1v_1v_2}} & q_{\mathtt{v_2v_1v_2v_1}} & q_{\mathtt{v_2v_1v_2v_2}} \\
q_{\mathtt{v_2v_2u_1u_1}} & q_{\mathtt{v_2v_2u_1u_2}} & q_{\mathtt{v_2v_2u_2u_1}} & q_{\mathtt{v_2v_2u_2u_2}} & q_{\mathtt{v_2v_2v_1v_1}} & q_{\mathtt{v_2v_2v_1v_2}} & q_{\mathtt{v_2v_2v_2v_1}} & q_{\mathtt{v_2v_2v_2v_2}}
\end{pmatrix}$$

$$M_2 = \begin{pmatrix}
q_{\mathtt{u_1v_1u_1v_1}} & q_{\mathtt{u_1v_1u_1v_2}} & q_{\mathtt{u_1v_1u_2v_1}} & q_{\mathtt{u_1v_1u_2v_2}} & q_{\mathtt{u_1v_1v_1u_1}} & q_{\mathtt{u_1v_1v_1u_2}} & q_{\mathtt{u_1v_1v_2u_1}} & q_{\mathtt{u_1v_1v_2u_2}} \\
q_{\mathtt{u_1v_2u_1v_1}} & q_{\mathtt{u_1v_2u_1v_2}} & q_{\mathtt{u_1v_2u_2v_1}} & q_{\mathtt{u_1v_2u_2v_2}} & q_{\mathtt{u_1v_2v_1u_1}} & q_{\mathtt{u_1v_2v_1u_2}} & q_{\mathtt{u_1v_2v_2u_1}} & q_{\mathtt{u_1v_2v_2u_2}} \\
q_{\mathtt{u_2v_1u_1v_1}} & q_{\mathtt{u_2v_1u_1v_2}} & q_{\mathtt{u_2v_1u_2v_1}} & q_{\mathtt{u_2v_1u_2v_2}} & q_{\mathtt{u_2v_1v_1u_1}} & q_{\mathtt{u_2v_1v_1u_2}} & q_{\mathtt{u_2v_1v_2u_1}} & q_{\mathtt{u_2v_1v_2u_2}} \\
q_{\mathtt{u_2v_2u_1v_1}} & q_{\mathtt{u_2v_2u_1v_2}} & q_{\mathtt{u_2v_2u_2v_1}} & q_{\mathtt{u_2v_2u_2v_2}} & q_{\mathtt{u_2v_2v_1u_1}} & q_{\mathtt{u_2v_2v_1u_2}} & q_{\mathtt{u_2v_2v_2u_1}} & q_{\mathtt{u_2v_2v_2u_2}} \\
q_{\mathtt{v_1u_1u_1v_1}} & q_{\mathtt{v_1u_1u_1v_2}} & q_{\mathtt{v_1u_1u_2v_1}} & q_{\mathtt{v_1u_1u_2v_2}} & q_{\mathtt{v_1u_1v_1u_1}} & q_{\mathtt{v_1u_1v_1u_2}} & q_{\mathtt{v_1u_1v_2u_1}} & q_{\mathtt{v_1u_1v_2u_2}} \\
q_{\mathtt{v_1u_2u_1v_1}} & q_{\mathtt{v_1u_2u_1v_2}} & q_{\mathtt{v_1u_2u_2v_1}} & q_{\mathtt{v_1u_2u_2v_2}} & q_{\mathtt{v_1u_2v_1u_1}} & q_{\mathtt{v_1u_2v_1u_2}} & q_{\mathtt{v_1u_2v_2u_1}} & q_{\mathtt{v_1u_2v_2u_2}} \\
q_{\mathtt{v_2u_1u_1v_1}} & q_{\mathtt{v_2u_1u_1v_2}} & q_{\mathtt{v_2u_1u_2v_1}} & q_{\mathtt{v_2u_1u_2v_2}} & q_{\mathtt{v_2u_1v_1u_1}} & q_{\mathtt{v_2u_1v_1u_2}} & q_{\mathtt{v_2u_1v_2u_1}} & q_{\mathtt{v_2u_1v_2u_2}} \\
q_{\mathtt{v_2u_2u_1v_1}} & q_{\mathtt{v_2u_2u_1v_2}} & q_{\mathtt{v_2u_2u_2v_1}} & q_{\mathtt{v_2u_2u_2v_2}} & q_{\mathtt{v_2u_2v_1u_1}} & q_{\mathtt{v_2u_2v_1u_2}} & q_{\mathtt{v_2u_2v_2u_1}} & q_{\mathtt{v_2u_2v_2u_2}}
\end{pmatrix}$$

and q_{xyzt} are the coordinates in the basis $x \otimes y \otimes z \otimes t$. We see that I_e is generated by $\binom{8}{3}\binom{8}{3} + \binom{8}{3}\binom{8}{3} = 6272$ polynomials of degree 3.

5.3. Kimura 3-parameter model

Take $G = \langle(\mathtt{AC})(\mathtt{GT}), (\mathtt{AG})(\mathtt{CT})\rangle$, which is also isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. The equivariant matrices for this group have the following structure:

$$\begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}$$

In this case, the equivariant model is the *Kimura 3-parameter model* introduced in [19]. We write $\omega_{\mathtt{A}}, \omega_{\mathtt{C}}, \omega_{\mathtt{G}}, \omega_{\mathtt{T}}$ for the irreducible characters of G. The corresponding table is

Ω_G	id	$(\mathtt{AC})(\mathtt{GT})$	$(\mathtt{AG})(\mathtt{CT})$	$(\mathtt{AT})(\mathtt{CG})$
$\omega_{\mathtt{A}}$	1	1	1	1
$\omega_{\mathtt{C}}$	1	1	-1	-1
$\omega_{\mathtt{G}}$	1	-1	1	-1
$\omega_{\mathtt{T}}$	1	-1	-1	1
χ	4	0	0	0

It follows that $\chi = \omega_{\mathtt{A}} + \omega_{\mathtt{C}} + \omega_{\mathtt{G}} + \omega_{\mathtt{T}}$ and so, $\mathbf{m} = (1,1,1,1)$

$$W = W[\omega_{\mathtt{A}}] \oplus W[\omega_{\mathtt{C}}] \oplus W[\omega_{\mathtt{G}}] \oplus W[\omega_{\mathtt{T}}],$$

where

$$W[\omega_{\mathtt{A}}] \cong N_{\omega_{\mathtt{A}}} \qquad W[\omega_{\mathtt{C}}] \cong N_{\omega_{\mathtt{C}}} \qquad W[\omega_{\mathtt{G}}] \cong N_{\omega_{\mathtt{G}}} \qquad W[\omega_{\mathtt{T}}] \cong N_{\omega_{\mathtt{T}}}.$$

In fact, if we write

$$\begin{aligned} \overline{\mathtt{A}} &= \mathtt{A}+\mathtt{C}+\mathtt{G}+\mathtt{T} \qquad & \overline{\mathtt{C}} &= \mathtt{A}+\mathtt{C}-\mathtt{G}-\mathtt{T} \\ \overline{\mathtt{G}} &= \mathtt{A}-\mathtt{C}+\mathtt{G}-\mathtt{T} \qquad & \overline{\mathtt{T}} &= \mathtt{A}-\mathtt{C}-\mathtt{G}+\mathtt{T} \end{aligned} \tag{5.1}$$

we have

$$W[\omega_{\mathtt{A}}] = \langle \overline{\mathtt{A}} \rangle \qquad W[\omega_{\mathtt{C}}] = \langle \overline{\mathtt{C}} \rangle \qquad W[\omega_{\mathtt{G}}] = \langle \overline{\mathtt{G}} \rangle \qquad W[\omega_{\mathtt{T}}] = \langle \overline{\mathtt{T}} \rangle$$

We remark that the basis $\{\overline{\mathtt{A}}, \overline{\mathtt{C}}, \overline{\mathtt{G}}, \overline{\mathtt{T}}\}$ is the image of $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ by the Fourier transform described in [8] or [9].

Since $\chi^2 = 4\omega_{\mathtt{A}} + 4\omega_{\mathtt{C}} + 4\omega_{\mathtt{G}} + 4\omega_{\mathtt{T}}$, we have $\mathbf{m}(2) = (4,4,4,4)$. In virtue of Remark 5.1,

$$\begin{aligned} W \otimes W[\omega_{\mathtt{A}}] &= \langle \overline{\mathtt{A}} \otimes \overline{\mathtt{A}}, \overline{\mathtt{C}} \otimes \overline{\mathtt{C}}, \overline{\mathtt{G}} \otimes \overline{\mathtt{G}}, \overline{\mathtt{T}} \otimes \overline{\mathtt{T}} \rangle \\ W \otimes W[\omega_{\mathtt{C}}] &= \langle \overline{\mathtt{A}} \otimes \overline{\mathtt{C}}, \overline{\mathtt{C}} \otimes \overline{\mathtt{A}}, \overline{\mathtt{G}} \otimes \overline{\mathtt{T}}, \overline{\mathtt{T}} \otimes \overline{\mathtt{G}} \rangle \\ W \otimes W[\omega_{\mathtt{G}}] &= \langle \overline{\mathtt{A}} \otimes \overline{\mathtt{G}}, \overline{\mathtt{C}} \otimes \overline{\mathtt{T}}, \overline{\mathtt{G}} \otimes \overline{\mathtt{A}}, \overline{\mathtt{T}} \otimes \overline{\mathtt{C}} \rangle \\ W \otimes W[\omega_{\mathtt{T}}] &= \langle \overline{\mathtt{A}} \otimes \overline{\mathtt{T}}, \overline{\mathtt{C}} \otimes \overline{\mathtt{G}}, \overline{\mathtt{G}} \otimes \overline{\mathtt{C}}, \overline{\mathtt{T}} \otimes \overline{\mathtt{A}} \rangle \end{aligned}$$

Then, I_e is given by the conditions

$$\mathbf{rk} \begin{pmatrix} M_{\mathtt{A}} & 0 & 0 & 0 \\ 0 & M_{\mathtt{C}} & 0 & 0 \\ 0 & 0 & M_{\mathtt{G}} & 0 \\ 0 & 0 & 0 & M_{\mathtt{T}} \end{pmatrix} \leq (1,1,1,1) \tag{5.2}$$

where $M_Z \in M_{4,4}$ for all $Z \in B$, that is,

$$M_{\mathtt{A}} = \begin{pmatrix} q_{\mathtt{AAAA}} & q_{\mathtt{AACC}} & q_{\mathtt{AAGG}} & q_{\mathtt{AATT}} \\ q_{\mathtt{CCAA}} & q_{\mathtt{CCCC}} & q_{\mathtt{CCGG}} & q_{\mathtt{CCTT}} \\ q_{\mathtt{GGAA}} & q_{\mathtt{GGCC}} & q_{\mathtt{GGGG}} & q_{\mathtt{GGTT}} \\ q_{\mathtt{TTAA}} & q_{\mathtt{TTCC}} & q_{\mathtt{TTGG}} & q_{\mathtt{TTTT}} \end{pmatrix} \quad M_{\mathtt{C}} = \begin{pmatrix} q_{\mathtt{ACAC}} & q_{\mathtt{AACA}} & q_{\mathtt{AAGT}} & q_{\mathtt{AATG}} \\ q_{\mathtt{CAAC}} & q_{\mathtt{CACA}} & q_{\mathtt{CAGT}} & q_{\mathtt{CATG}} \\ q_{\mathtt{GTAC}} & q_{\mathtt{GTCA}} & q_{\mathtt{GTGT}} & q_{\mathtt{GTTG}} \\ q_{\mathtt{TGAC}} & q_{\mathtt{TGCA}} & q_{\mathtt{TGGT}} & q_{\mathtt{TGTG}} \end{pmatrix}$$

$$M_{\mathtt{G}} = \begin{pmatrix} q_{\mathtt{AGAG}} & q_{\mathtt{AGCT}} & q_{\mathtt{AGGA}} & q_{\mathtt{AGTC}} \\ q_{\mathtt{CTAG}} & q_{\mathtt{CTCT}} & q_{\mathtt{CTGA}} & q_{\mathtt{CTTC}} \\ q_{\mathtt{GAAG}} & q_{\mathtt{GACT}} & q_{\mathtt{GAGA}} & q_{\mathtt{GATC}} \\ q_{\mathtt{TCAG}} & q_{\mathtt{TCCT}} & q_{\mathtt{TCGA}} & q_{\mathtt{TCTC}} \end{pmatrix} \quad M_{\mathtt{T}} = \begin{pmatrix} q_{\mathtt{ATAT}} & q_{\mathtt{ATCG}} & q_{\mathtt{ATGC}} & q_{\mathtt{ATTA}} \\ q_{\mathtt{CGAT}} & q_{\mathtt{CGCG}} & q_{\mathtt{CGGC}} & q_{\mathtt{CGTA}} \\ q_{\mathtt{GCAT}} & q_{\mathtt{GCCG}} & q_{\mathtt{GCGC}} & q_{\mathtt{GCTA}} \\ q_{\mathtt{TAAT}} & q_{\mathtt{TACG}} & q_{\mathtt{TAGC}} & q_{\mathtt{TATA}} \end{pmatrix}$$

where $q_{\mathtt{X_1X_2X_3X_4}}$ are the coordinates in the basis $\{\overline{\mathtt{X}}_1 \otimes \overline{\mathtt{X}}_2 \otimes \overline{\mathtt{X}}_3 \otimes \overline{\mathtt{X}}_4\}_{\mathtt{X_i} \in \mathtt{B}}$. The ideal I_e obtained by imposing the rank conditions of (5.2) is generated by $\binom{4}{2}\binom{4}{2} + \binom{4}{2}\binom{4}{2} + \binom{4}{2}\binom{4}{2} + \binom{4}{2}\binom{4}{2} = 144$ quadrics. However, at any point of $V(I_e)$ the variety is locally defined by 36 quadrics (see [8, Example 4.9]).

5.4. Kimura 2-parameter model

Take $G = \langle(\texttt{ACGT}), (\texttt{AG})\rangle$, which is isomorphic to the dihedral group. The equivariant matrices for this group have the following structure:

$$\begin{pmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{pmatrix}$$

The equivariant model is the *Kimura 2-parameter model* introduced in [18]. There are 5 irreducible characters ω_1, ω_2, ω_3, ω_4, ω and the corresponding table is

Ω_G	id	$(\texttt{ACGT})$	$(\texttt{AG})$	$(\texttt{AG})(\texttt{CT})$	$(\texttt{AC})(\texttt{GT})$
ω_1	1	1	1	1	1
ω_2	1	1	-1	1	-1
ω_3	1	-1	1	1	-1
ω_4	1	-1	-1	1	1
ω	2	0	0	-2	0
χ	4	0	2	0	0

Notice that G is not abelian and that the irreducible representation ω is 2-dimensional. It follows that $\chi = \omega_1 + \omega_3 + \omega$ and so, $\mathbf{m} = (1, 0, 1, 0, 1)$ and

$$W = W[\omega_1] \oplus W[\omega_3] \oplus W[\omega],$$

where

$$W[\omega_1] \cong N_{\omega_1} \qquad W[\omega_3] \cong N_{\omega_3} \qquad W[\omega] \cong N_{\omega}.$$

In fact, with the notation of (5.1) we have

$$W[\omega_1] = \langle\overline{\texttt{A}}\rangle \qquad W[\omega_3] = \langle\overline{\texttt{G}}\rangle \qquad W[\omega] = \langle\overline{\texttt{C}}, \overline{\texttt{T}}\rangle$$

Now, we consider the case of four leaves. We have $\chi^2 = 3\omega_1 + \omega_2 + 3\omega_3 + \omega_4 + 4\omega$, so $\mathbf{m}(2) = (3, 1, 3, 1, 4)$. If $\psi \in \mathcal{L}(T)^G$, then

$$Tf_e(\psi) = \begin{pmatrix} S_1 & 0 & 0 & 0 & 0 \\ 0 & S_2 & 0 & 0 & 0 \\ 0 & 0 & S_3 & 0 & 0 \\ 0 & 0 & 0 & S_4 & 0 \\ 0 & 0 & 0 & 0 & S \end{pmatrix} \in M_{\mathbf{m}(2),\mathbf{m}(2)}$$

where

$$S_1 \in M_{3,3} \qquad S_2 \in M_{1,1} \qquad S_3 \in M_{3,3} \qquad S_4 \in M_{1,1} \qquad S \in M_{4,4}.$$

Then, the ideal I_e is given by the condition

$$\mathbf{rk}\ Tf_{L_1|L_2}(\psi) \leq (1,0,1,0,1).$$

By imposing these rank conditions to the matrix $Tf_{L_1,L_2}(\psi)$ we obtain $\binom{3}{2}\binom{3}{2} + \binom{1}{1}\binom{1}{1} + \binom{3}{2}\binom{3}{2} + \binom{1}{1}\binom{1}{1} + \binom{4}{2}\binom{4}{2} = 9+1+9+1+36 = 56$ invariants: 54 of them are quadrics and 2 of them are linear invariants.

5.5. *Jukes-Cantor model*

Finally, we take the whole group of permutations $\mathfrak{S}_4$. The equivariant matrices for this group have the following structure:

$$\begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}$$

The equivariant model associated to it is the *Jukes-Cantor model* introduced in [17]. The group $\mathfrak{S}_4$ has five irreducible characters $\{\omega_i\}_{i=0,\dots,4}$ (see §2.3 of [15]) and the following character table:

$\Omega_{\mathfrak{S}_4}$	id	(AC)	(ACG)	(ACGT)	(AC)(GT)
ω_0	1	1	1	1	1
ω_1	1	-1	1	-1	1
ω_2	2	0	-1	0	2
ω_3	3	1	0	-1	-1
ω_4	3	-1	0	1	- 1
χ	4	2	1	0	0

It follows that

$$\chi = \omega_0 + \omega_3,$$

that is, χ is the sum of the trivial and the *standard* representations. We have $\mathbf{m} = (1,0,0,1,0)$. Thus, there is a decomposition

$$W = W[\omega_0] \oplus W[\omega_3],$$

where

$$\begin{aligned} W[\omega_0] &\cong N_{\omega_0} \otimes \mathbb{C}^{m_0} \cong N_{\omega_0} \qquad && \dim W[\omega_0] = 1 \\ W[\omega_3] &\cong N_{\omega_3} \otimes \mathbb{C}^{m_3} \cong N_{\omega_3} \qquad && \dim W[\omega_3] = 3. \end{aligned}$$

In fact, with the notation of (5.1), we have

$$W[\omega_0] = \langle \overline{\mathtt{A}} \rangle \qquad W[\omega_3] = \langle \overline{\mathtt{C}}, \overline{\mathtt{G}}, \overline{\mathtt{T}} \rangle.$$

The ideal I_e is generated by the $(m_j + 1)$-minors of the j-th box of $Tf_e(\psi)$ with $j = 1, 2, \ldots, 5$. On the other hand, it is straightforward to see that $\chi^2 = 2\omega_0 + \omega_2 + 3\omega_3 + \omega_4$, so $\mathbf{m}(2) = (2, 0, 1, 3, 1)$ and we have

$$\begin{array}{rcl}
(W \otimes W)[\omega_0] & = & \langle q_{\mathtt{AA}}, q_{\mathtt{CC}} + q_{\mathtt{GG}} + q_{\mathtt{TT}} \rangle \\
(W \otimes W)[\omega_2] & = & \langle q_{\mathtt{CC}} - q_{\mathtt{GG}}, q_{\mathtt{CC}} - q_{\mathtt{TT}} \rangle \\
(W \otimes W)[\omega_3] & = & \langle q_{\mathtt{AC}}, q_{\mathtt{AG}}, q_{\mathtt{AT}}, q_{\mathtt{CA}}, q_{\mathtt{GA}}, q_{\mathtt{TA}}, q_{\mathtt{CT}} + q_{\mathtt{TC}}, q_{\mathtt{CG}} + q_{\mathtt{GC}}, q_{\mathtt{GT}} + q_{\mathtt{TG}} \rangle \\
(W \otimes W)[\omega_4] & = & \langle q_{\mathtt{CT}} - q_{\mathtt{TC}}, q_{\mathtt{CG}} - q_{\mathtt{GC}}, q_{\mathtt{GT}} - q_{\mathtt{TG}} \rangle
\end{array}$$

and $q_{\mathtt{XY}} = q_{\mathtt{X}} \otimes q_{\mathtt{Y}}$, for any $\mathtt{X}, \mathtt{Y} \in B$. Now, if $\psi \in \mathcal{L}(T)^{\mathfrak{S}_4}$ we have

$$Tf_e(\psi) = \begin{pmatrix} S_0 & 0 & 0 & 0 \\ 0 & S_2 & 0 & 0 \\ 0 & 0 & S_3 & 0 \\ 0 & 0 & 0 & S_4 \end{pmatrix} \in M_{\mathbf{m}(2),\mathbf{m}(2)}$$

where

$$S_0 \in M_{2,2} \qquad S_2 \in M_{1,1} \qquad S_3 \in M_{3,3} \qquad S_4 \in M_{1,1}.$$

For instance, we have

$$S_0 = \left(\begin{array}{c|c} q_{\mathtt{AAAA}} & q_{\mathtt{AACC}} + q_{\mathtt{AAGG}} + q_{\mathtt{AATT}} \\ \hline q_{\mathtt{CCAA}} + q_{\mathtt{GGAA}} + q_{\mathtt{TTAA}} & \begin{smallmatrix} q_{\mathtt{CCCC}} + q_{\mathtt{GGCC}} + q_{\mathtt{TTCC}} + \\ q_{\mathtt{CCGG}} + q_{\mathtt{GGGG}} + q_{\mathtt{TTGG}} + \\ q_{\mathtt{CCTT}} + q_{\mathtt{GGTT}} + q_{\mathtt{TTTT}} \end{smallmatrix} \end{array} \right)$$

while

$$S_2 = (q_{\mathtt{CCCC}} - q_{\mathtt{CCGG}} - q_{\mathtt{GGCC}} + q_{\mathtt{GGGG}}).$$

Now, given $\psi \in \mathcal{L}(T)^{\mathfrak{S}_4}$, we have $\psi \in V(T)$ if and only if

$$\mathbf{rk}\ Tf_e(\psi) \leq \mathbf{m}. \tag{5.3}$$

By imposing these rank conditions to the matrix $Tf_e(\psi)$ we obtain $\binom{2}{2}\binom{2}{2} + 0 + \binom{1}{1}\binom{1}{1} + \binom{3}{2}\binom{3}{2} + \binom{1}{1}\binom{1}{1} = 12$ phylogenetic invariants $\{f_i\}_{i=1,\ldots,12}$:

1. $f_1, \ldots, f_{10}$ have degree 2 and are obtained by the conditions $\mathrm{rk}\,(S_0), \mathrm{rk}\,(S_3) = 1$

2. f_{11}, f_{12} have degree one and are obtained by the conditions $S_1, S_4 = 0$. These two invariants are equivalent to Lake's invariants (cf. [20]).

References

[1] E. Allman, J. Rhodes, Phylogenetic invariants for the general Markov model of sequence mutation, Math. Biosci. 186 (2003) 113–144.

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

[3] E. Allman, J. Rhodes, Phylogenetic ideals and varieties for the general Markov model, Adv. in Appl. Math. 40 (2007) 127–148.

[4] E. Allman, J. Rhodes, Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites, Mathematical Biosciences 211 (2008) 18–33.

[5] D. Barry, J. Hartigan, Asynchronous distance between homologous DNA sequences, Biometrics 43 (1987) 261–276.

[6] P. Buneman, The recovery of trees from measures of dissimilarity, in: E.U. Press (Ed.), Mathematics in the Archaeological and Historical Sciences, pp. 387–395.

[7] M. Casanellas, J. Fernandez-Sanchez, Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees, Mol. Biol. Evol. 24 (2007) 288–293.

[8] M. Casanellas, J. Fernandez-Sanchez, Geometry of the kimura 3-parameter model, Adv. in Appl. Math 41 (2008) 265–292.

[9] M. Casanellas, L. Garcia, S. Sullivant, Catalog of small trees, in: L. Pachter, B. Sturmfels (Eds.), Algebraic Statistics for computational biology, Cambridge University Press, 2005.

[10] M. Casanellas, S. Sullivant, The strand symmetric model, in: L. Pachter, B. Sturmfels (Eds.), Algebraic Statistics for computational biology, Cambridge University Press, 2005.

[11] J. Cavender, J. Felsenstein, Invariants of phylogenies in a simple case with discrete states, J. Classification 4 (1987) 57–71.

[12] J. Draisma, J. Kuttler, On the ideals of equivariants tree models, Mathematische Annalen 344 (2009) 619–644.

[13] N. Eriksson, Tree construction using singular value decomposition, in: L. Pachter, B. Sturmfels (Eds.), Algebraic Statistics for computational biology, Cambridge University Press, 2005, pp. 347–358.

[14] J. Felsenstein, Counting phylogenetic invariants in some simple cases, Journal of Theoretical Biology 152 (1991) 357 376.

[15] W. Fulton, J. Harris, Representation theory, volume 129 of *Graduate Texts in Mathematics*, Springer-Verlag, New York, 1991. A first course, Readings in Mathematics.

[16] L. Garcia-Puente, Algebraic statistics in model selection, in: M. Chickering, J. Halpern (Eds.), Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 177–184.

[17] T. Jukes, C. Cantor, Evolution of protein molecules., In Mammalian Protein Metabolism (1969) 21–132.

[18] M. Kimura, A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences, J. Mol. Evol. 16 (1980) 111–120.

[19] M. Kimura, Estimation of evolutionary sequences between homologous nucleotide sequences, Proc. Nat. Acad. Sci. , USA 78 (1981) 454–458.

[20] J. Lake, A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony, Mol. Biol. Evol. 4 (1987) 167–191.

[21] L. Pachter, B. Sturmfels (Eds.), Algebraic Statistics for computational biology, Cambride University Press, 2005. ISBN 0-521-85700-7.

[22] J. 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.

[23] M. Steel, Recovering a tree from the leaf colourations it generates under a markov model, Applied Mathematics Letters 7 (1994) 19–24.

[24] M. Steel, M. Hendy, L. Székely, P. Erdős, Spectral analysis and a closest tree method for genetic sequences, Appl. Math. Lett. 5 (1992) 63–67.

[25] B. Sturmfels, S. Sullivant, Toric ideals of phylogenetic invariants, J. Comput. Biol. 12 (2005) 204–228.