REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Stochasticity conditions for the general Markov model

*Marina Garrote López
Universitat Politècnica de Catalunya
marinagarrotelopez@gmail.com

*Corresponding author

Resum *(CAT)*

En filogenètica sovint es modelitza l'evolució mol·lecular mitjanant models estadístics paramètrics. Usant aquests models es poden deduir relacions polinomials (*invariants filogenètics*) entre les probabilitats teòriques de caràcters observats a les fulles de l'arbre. En aquest article estudiem i corregim alguns resultats teòrics que ens proporcionaran condicions en l'estocasticitat dels paràmetres d'aquests models i que utilitzem per tal de trobar nous invariants filogenètics.

Abstract *(ENG)*

In phylogenetics it is useful to model evolution adopting parametric statistic models. Using these models one is able to deduce polynomial relationships between the theoretical probabilities of characters at the leaves of a phylogenetic tree, known as *phylogenetic invariants*. We revisit and correct some results on stochasticity conditions of the parameters of these models and we find new phylogenetic invariants.

Keywords: *Phylogenetic tree, phylogenetic invariants, topology invariants, general Markov model, joint distribution, tensor.*
MSC (2010): *92D15, 92D20, 14P10, 60J20, 62P10.*
Received: *February 3th, 2016.*
Accepted: *March 9th, 2016.*

Acknowledgement

I would like to thank my advisors Marta Casanellas and Jesús Fernández-Sánchez for having invested a lot of their time in this research. I am very grateful to them for sharing with me their scientific knowledge and for their unconditional support.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

1. Introduction

Strong evidences suggest that all living organisms share a common ancestor and therefore, are related by evolutionary relationships. These relationships are usually expressed in the form of a phylogenetic tree.

Nowadays there are more and more mathematicians and statisticians who collaborate with biologists in order to solve the major problems of phylogenetics. Many different areas of mathematics, like statistics, probability, algebra, combinatorics and numerical methods are involved in phylogenetic studies. Even more, recently developed techniques from algebraic geometry have already been used in the study of phylogenetics.

The main goal of phylogenetic reconstruction is recovering the ancestral relationships among a group of current species. In order to reconstruct phylogenetic trees it is necessary to model evolution adopting a parametric statistic model. Using these models one is able to deduce polynomial relationships between the parameters of the model, known as *phylogenetic invariants*. Mathematicians have recently begun to be interested in the study of these polynomials and the geometry of the algebraic varieties that arise in this setting. Furthermore, they have started to use some phylogenetic invariants called *topology invariants* to reconstruct phylogenetic trees; see [4, 8].

The aim of this paper is to understand the relationship between phylogenetics and these algebraic techniques to recover phylogenetic trees from real data. Our main goal is to study and to analyze the characterizations of stochasticity of the points in the algebraic varieties mentioned above, and provided in [5].

The paper is divided into two parts. In the first one, we explain basic concepts on phylogenetics that are already known. We explain what *phylogenetic trees* are from the mathematical standpoint, we describe the general Markov model, and we explain then what *phylogenetic invariants* and *topology invariants* are. Moreover, we define *joint distributions* of a tree and its representation as a tensor. We will define some operations among tensors that will be useful, and their meaning in terms of phylogenetic trees. This part will be developed in Section 2. After that, in Section 3, we will revisit results related to the stochasticity of the parameters of the general Markov model on a tree. One of these results, [5, Theorem 3.2.4], has been restated and the proof rewritten since the statement of the original theorem is not completely correct. We also provide a counterexample to show this; see Counterexample 3.8. Finally, in Theorem 3.11 we present new topology invariants that can be used to design original methods for phylogenetic reconstruction; see [10] for further details.

2. Preliminaries

2.1 Biological preliminaries

Phylogenetics is the study of relationships between different species or biological entities. It studies how species evolve and where contemporary species come from. According to the theory of the biological evolution developed by Darwin (s.XIX), all species of organisms evolve through the natural selection of small variations that increase the individual's ability to compete, survive, and reproduce. We can model these specialization processes with phylogenetic trees. The nodes of these trees represent different species and every branch is an evolutionary process between two species. The leaves of the tree are contemporary species and the root of the tree is the common ancestor of all the species represented on the tree.

Genetic information of each individual is encoded in the DNA of the nucleus of its cells, which is composed of four different simpler units named *nucleotides*. According to the bases forming the nucleotides, they are called adenine (A), cytosine (C), guanine (G) and thymine (T).

Heredity information in a genome is thought to be contained in genes. But DNA sequences of a same gene may look quite different for different species. They contain similar parts but they can also contain some other parts that can not be compared. For that reason the first problem is identifying which part of DNA sequences of different species can be compared. This information is collected in an *alignment*. A sequence alignment is a way of arranging the sequences of DNA to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the species. We can represent an alignment with a table whose rows are DNA sequences of the species and whose columns correspond to nucleotides evolved from the same nucleotide at the common ancestor of all the species in the table. Alignments are used in many contexts, phylogenetics among them, to see relationships between some species and to reconstruct the phylogenetic tree relating them.

One of the basic objects in a phylogenetic model is a tree T encoding the evolutionary relationships among a given set of species. In this section we introduce some concepts that allow us to deal with these phylogenetic trees following the approach in [2, 3, 6].

Definition 2.1. A *tree* T is a connected graph with no cycles. The *degree* of a vertex is the number of edges incident to it. The vertices of degree 1 are called *leaves* and the set of leaves of T is denoted by $L(T)$. All the other vertices, which have degree at least 2, are *interior nodes* and are designated by the set $Int(T)$. $E(T)$ is the set of the edges of the tree. If all nodes in $Int(T)$ have degree 3, then T is called a *trivalent tree*. A tree is called a *rooted tree* if one vertex has been labelled as "root", and the edges are oriented away from it. A *phylogenetic tree* is a pair (T, ϕ), where T is a tree and $\phi\colon X \to L(T)$ is a one-to-one correspondence between the set of leaves and a finite set of labels denoted be X. The *tree topology* of a phylogenetic tree is the topology of the tree as a labelled graph.

In a phylogenetic tree, the set X represents a set of living species and the tree T shows the ancestral relationships among them. Every edge represents an evolutionary process between two species and if it is rooted, then the root represents the common ancestor to the set of species X. For our purposes, usually X will be taken as the set $\{1, 2, \dots, n\}$. Moreover, two phylogenetic trees T_1 and T_2, with the same set of labels X at the leaves, have the same topology if there is a one-to-one correspondence φ between their vertices respecting adjacency and leaf labelling. If r_1, r_2 are the roots of T_1 and T_2, respectively, then we need to impose $\varphi(r_1) = r_2$.

Remark 2.2. For the rest of the paper, we denote by T_n the set of all possible tree topologies for n-leaf unrooted trivalent trees. Note that n has to be greater or equal than 3 ($|T_3| = 1$). We will denote the three possible topologies of T_4 by $T_{12|34}$, $T_{13|24}$, and $T_{14|23}$; see left hand side of Figure 1.

2.2 Evolutionary models

Evolution is usually modeled adopting a parametric statistical model. That is, evolution is assumed to be a stochastic process, in which nucleotides mutate randomly over time according to certain probabilities. Moreover we assume that DNA substitutions occur randomly and the nucleotides observed in the DNA sequences are independent and identically distributed.

We associate a discrete random variable X_i to each node i of T such that X_i can take κ different states. We denote by $\mathcal{K}$ this set of states. Usually $\mathcal{K}$ is the set of the four nucleotides in DNA, which are

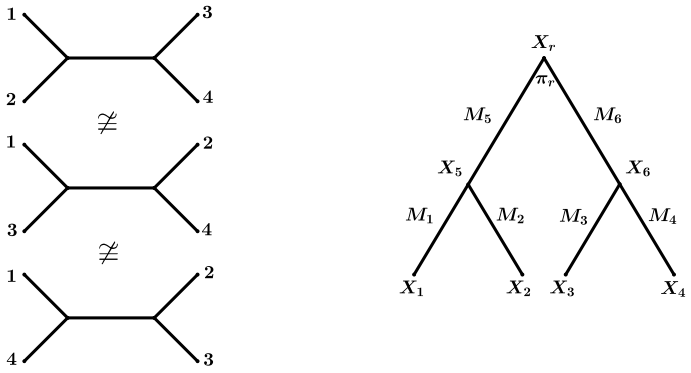

Figure 1: *Left:* the three topologies of T_4, say $T_{12|34}$, $T_{13|24}$, and $T_{14|23}$. *Right:* a Markov process on a rooted 4-leaf tree given by a distribution vector π and transition matrices $\{M_1, \dots, M_6\}$.

denoted by their first letter, so $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and $\kappa = 4$. Since DNA sequences of the contemporary species are known, we say that random variables at the leaves are observed. However, we do not have any information about the ancestral species, that is why random variables at the interior nodes are hidden. For a tree T with leaves $1, 2, \dots, n$, $X = (X_1, X_2, \dots, X_n)$ represents the joint distribution vector of the leaves. Each column of an alignment is an observation of this vector of random variables.

Hereafter we introduce a Markov process in a rooted tree T. First, we define a vector $\pi = (\pi_1, \dots, \pi_\kappa)$, the distribution of X_r which is the random variable associated to the root r and satisfying that all entries are nonnegative and $\sum_i \pi_i = 1$. If $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, we interpret these entries as the probabilities that an arbitrary site in the DNA sequence at the root is occupied by the corresponding base. A second set of parameters is associated to the evolutionary process that occurs in every edge. For each edge e we associate a $\kappa \times \kappa$ matrix M_e, called *substitution* or *transition* matrix.

Definition 2.3. A *transition matrix* is a $\kappa \times \kappa$ matrix M_e associated to each edge of a phylogenetic tree. Every entry is the conditional probability $P(x|y, e)$ that a state y at the parent node of e had been substituted by a state x at its child, during the evolutionary process along the edge e. Since each row contains the probabilities of the κ possible changes that can occur in an evolutionary process, rows of M_e sum up to 1. These matrices M_e are also called *Markov matrices* or *row stochastic matrices*.

We consider a Markov process on T given by π and the matrices $\{M_E\}_{e \in E(T)}$. In particular, the substitutions on two adjacent branches at a node v are independent given the state at v.

The substitution probabilities on a given edge depend only on the state at the parent node. Besides, we only have observations of the random variables at the leaves so, ours is a *hidden* Markov process. According to the shape of the transition matrices different *models* are defined, but in this paper we focus on the general Markov model, that is, transition matrices do not satisfy any other restriction.

Example 2.4. On the right hand side of Figure 1, a Markov process on a phylogenetic tree is represented. The $X_i's$ are random variables associated to the leaves, the $M_i's$ are the transition matrices, and π_r is the root distribution. Under the general Markov model, we have 3×4 free parameters for each transition matrix and 3 free parameters for the vector π_r. Therefore, this model has $3 \cdot 4 \cdot 6 + 3 = 75$ free parameters.

In what follows we describe how to compute the joint probability of observing states $x_1, x_2, \dots, x_n$ at the leaves according to the Markov process we have described.

We denote by $p_{x_1,\dots,x_n}$ the joint distribution at the leaves of a rooted phylogenetic tree T, $p_{x_1,\dots,x_n} = \text{Prob}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$. We define P as a κ^n-dimensional vector whose components are the joint probabilities $p_{x_1,\dots,x_n}$, $P = (p_{x_1,\dots,x_n})_{x_1,\dots,x_n \in \mathcal{K}}$.

Since the evolutionary processes follow a Markov process, they are independent and we can express $p_{x_1,\dots,x_n}$ in terms of the transition matrices,

$$p_{x_1,\dots,x_n} = \sum_{x_r,(x_v)_{v\in \text{Int}(T)}} \prod_{e\in E(T)} M_e(x_{a(e)}, x_{d(e)}), \tag{1}$$

where $x_r \in \mathcal{K}$ is a state of the root, $x_{a(e)} \in \mathcal{K}$ is a state of the parent node of the edge e, and $x_{d(e)} \in \mathcal{K}$ is the state of the descendant node of the edge e. If e is a terminal edge ending at the leaf i then $x_{d(e)} = x_i$. Every entry of P can be seen as a polynomial with the parameters of the model $\mathcal{M}$ as variables.

Example 2.5. We compute now the joint distribution p_{x_1,x_2,x_3,x_4} of the tree presented on the right hand side of Figure 1. Using equation (1) we get

$$p_{x_1,x_2,x_3,x_4} = \sum_{x_r\in\mathcal{K}} \sum_{x_5\in\mathcal{K}} \sum_{x_6\in\mathcal{K}} \pi_{x_r} \cdot M_5(x_r, x_5) \cdot M_1(x_5, x_1) \cdot M_2(x_5, x_2) \cdot M_6(x_r, x_6) \cdot M_3(x_6, x_3) \cdot M_4(x_6, x_4).$$

2.3 Phylogenetic invariants and flattening

It is known that there exist many algebraic relations among the components of the joint distribution P; see [4, 6, 7, 9].

Since components of P are polynomials in the model parameters, we can associate to the tree a polynomial map $\varphi_T \colon \mathbb{C}^d \to \mathbb{C}^{\kappa^n}$ mapping any d-tuple of parameters to a distribution vector of the κ^n possible observations at the leaves of T. More precisely, we define the map

$$\begin{array}{rcl} \varphi_T \colon \mathbb{C}^d & \longrightarrow & \mathbb{C}^{\kappa^n} \\ (\pi, \{M_e\}_{e\in E(T)}) & \mapsto & P = (p_{x_1,x_1,\dots,x_1}, p_{x_1,x_1,\dots,x_2}, p_{x_1,x_1,\dots,x_3}, \dots, p_{x_\kappa,x_\kappa,\dots,x_\kappa}), \end{array} \tag{2}$$

where d is the number of free parameters of the model and each component $p_{x_1,\dots,x_n}$ is expressed in terms of the root distribution π and the transition matrices M_e according to the expression (1).

Remark 2.6. Notice that, to read the parameters as probabilities, we should restrict to nonnegative real numbers. Analogously, the points in the image of φ_T represent a joint distribution only if they lie in the standard $(\kappa^n - 1)$-simplex. However, in order to use techniques from algebraic geometry, we abandon temporally these restrictions and work over the complex field. We will consider *complex parameters* and complex parametrization map in general, but we will refer to *stochastic parameters* to the ones coming from the original probabilistic model (that is, all the components of π and the entries of the transition matrices M_i are nonnegative).

We introduce now an algebraic variety in $\mathbb{C}^{\kappa^n}$ which contains the set of image points of φ_T.

Definition 2.7. The *phylogenetic variety* associated to a tree T, denoted by $\mathcal{V}(T)$, is the smallest algebraic variety containing the image $\text{Im}\,\varphi_T$.

Remark 2.8. The image set $\text{Im}\,\varphi_T$ is not, in general, an algebraic variety, but it defines a dense open subset in $\mathcal{V}(T)$ under Zariski topology. The ideal $I(\text{Im}\,\varphi_T)$ of all polynomial relations in $\mathbb{C}[P_{x_1,\dots,x_n}]$ of the points in $\text{Im}(\varphi_T)$ coincides with the ideal of the variety $\mathcal{V}(T)$. We will denote it by $I(T)$. It can be proved that $\mathcal{V}(T)$ is independent from the node chosen as root in T; see [1] for a complete proof.

Definition 2.9. The polynomials in $I(T)$ are called *phylogenetic invariants* of T. If f is a polynomial in $I_{\mathcal{M}}(T)$ that does not belong to $I(T')$ for some other tree topology T' on n leaves, then f is called a *topology invariant* of T.

Definition 2.10. Let $A|B$ be a partition of the leaves of a tree T, that is $A, B \subseteq L(T)$, with $|A|, |B| \geq 2$ such that $L(T) = A \cup B$ and $A \cap B = \emptyset$. Let $\tilde{X}_A = (x_i)_{i\in A}$ and $\tilde{X}_B = (x_j)_{j\in B}$ be the random variables associated to A and B. Then $\tilde{X}_A$ and $\tilde{X}_B$ can take $a := \kappa^{|A|}$ and $b := \kappa^{|B|}$ states, respectively. Given a vector $P \in \mathbb{C}^{\kappa^n}$ we define the *flattening* $Flatt_{A|B}(P)$ as the $a \times b$ matrix whose entries are the joint distributions of all possible observations of $\tilde{X}_A$ and $\tilde{X}_B$:

$$Flatt_{A|B}(P) = \begin{array}{c} \\ \text{States of } \tilde{X}_A \end{array} \begin{array}{c} \text{States of } \tilde{X}_B \\ \begin{pmatrix} p_{u_1v_1} & p_{u_1v_2} & \cdots & p_{u_1v_b} \\ p_{u_2v_1} & p_{u_2v_2} & \cdots & p_{u_2v_b} \\ \vdots & \vdots & \ddots & \vdots \\ p_{u_av_1} & p_{u_av_2} & \cdots & p_{u_av_b} \end{pmatrix} \end{array}.$$

This matrix allows us to state the following result, which gives us some topology invariants associated to a 4-leaf tree.

Theorem 2.11 (Casanellas–Fernández-Sánchez, [8])**.** *Let T be a tree, $A|B$ a bipartition of $L(T)$ and $P = \varphi_T(\pi, \{M_e\}_{e\in E(T)})$. Then the $(\kappa+1)\times(\kappa+1)$ minors of $Flatt_{A|B}(P)$ vanish if $A|B$ is induced by removing an edge of T. Otherwise, $Flatt_{A|B}(P)$ has rank $\geq \kappa^2$ for general P. Therefore, the $(\kappa+1)\times(\kappa+1)$ minors of $Flatt_{A|B}(P)$ are topology invariants for the tree T.*

There is a more algebraic way of viewing the joint distribution at the leaves of a phylogenetic tree, which will be really useful in this article.

Let $\mathcal{W} := \mathbb{C}^{\kappa}$ be regarded as a vector space. We identify the canonical basis of $\mathcal{W}$ with the set $\mathcal{K}$. Then, the natural basis of $\mathcal{W} \otimes \overset{n)}{\cdots} \otimes \mathcal{W}$ is $\{x_1 \otimes \cdots \otimes x_n\}_{x_1,\ldots,x_n\in\mathcal{K}}$. For instance, if $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, the natural basis of $\mathcal{W} \otimes \mathcal{W} \otimes \mathcal{W}$ is $\{\mathtt{A}\otimes\mathtt{A}\otimes\mathtt{A},\ \mathtt{A}\otimes\mathtt{A}\otimes\mathtt{C},\ \ldots,\ \mathtt{T}\otimes\mathtt{T}\otimes\mathtt{T}\}$. Back to the description of the joint distribution $P = (p_{x_1,\ldots,x_n})_{x_1,\ldots,x_n\in\mathcal{K}}$ in the phylogenetic framework, we can think of P as a n-tensor in $\mathcal{W} \otimes \overset{n)}{\cdots} \otimes \mathcal{W}$ whose components in the natural basis above are $P = (p_{x_1,\ldots,x_n})_{x_1,\ldots,x_n\in\mathcal{K}}$:

$$P = \sum_{x_1,\ldots,x_n\in\mathcal{K}} p_{x_1,\ldots,x_n} x_1 \otimes \cdots \otimes x_n.$$

Each factor in $\mathcal{W} \otimes \overset{n)}{\cdots} \otimes \mathcal{W}$ corresponds to one specie so, in order to make species apparent in this tensor product, we denote it as $\mathcal{W}_1 \otimes \cdots \otimes \mathcal{W}_n$, where $\mathcal{W}_i = \mathcal{W}$ for every $i = 1, \ldots, n$. If we view the vector of joint distribution P as a tensor in $\mathcal{W}_1 \otimes \cdots \otimes \mathcal{W}_n$ then, keeping the notation of Definition 2.10, the flattening $Flatt_{A|B}(P)$ is the image of P via the isomorphism

$$\begin{array}{ccccc} \mathcal{W}_1 \otimes \cdots \otimes \mathcal{W}_n & \cong & Hom(\bigotimes_{i\in A} \mathcal{W}_i, \bigotimes_{j\in B} \mathcal{W}_j) & \cong & M_{a\times b}(\mathbb{C}), \\ P & & \longmapsto & & Flatt_{A|B}(P) \end{array}$$

where $M_{a\times b}(\mathbb{C})$ is the space of all $a \times b$ matrices with complex entries.

Notation 2.12. For the rest of the paper, given a vector $\mathbf{v} \in \mathbb{C}^\kappa$, $\mathbf{v}(i)$ will be the i-th component of $\mathbf{v}$ relative to the canonical basis $\{\mathbf{e_1}, \dots, \mathbf{e_\kappa}\}$ of $\mathbb{C}^\kappa$, and we will write $\mathbf{1}$ for $(1, \dots, 1)$. Moreover, we will call an *n-tensor* to the tensors $P \in \mathbb{C}^\kappa \otimes \overset{n)}{\cdots} \otimes \mathbb{C}^\kappa$, and it will be convenient to write $P(x_1, \dots, x_n)$ for the component $p_{x_1,\dots,x_n}$.

Definition 2.13. Given an n-tensor P, an integer $i \in \{1, \dots, n\}$ and a vector $\mathbf{v} \in \mathbb{C}^\kappa$, we define $P *_i \mathbf{v}$ the $(n-1)$-tensor given by $(P *_i \mathbf{v})(j_1, \dots, j_{i-1}, j_{i+1}, \dots, j_n) = \sum_{j_i=1}^{\kappa} v(j_i) P(j_1, \dots, j_i, \dots, j_n)$. We also define the *l-th slice of P in the i-th index* by $P_{\dots l \dots} = P *_i \mathbf{e_l}$, The *i-th marginalization of P* is defined as $P_{\dots + \dots} = P *_i \mathbf{1}$. Given a $\kappa \times \kappa$ matrix M, we define the *n-tensor* $P *_i M$ by

$$(P *_i M)(j_1, \dots, j_n) = \sum_{l=1}^{\kappa} P(j_1, \dots, j_{i-1}, l, j_{i+1}, \dots, j_n) M(l, j_i). \tag{3}$$

Remark 2.14. From now on, we consider the 2-tensors as $\kappa \times \kappa$ matrices via the isomorphism

$$P = \sum P(j_1, j_2) \mathbf{e_{j_1}} \otimes \mathbf{e_{j_2}} \leftrightarrow (P(j_1, j_2))_{j_1, j_2},$$

where rows of the matrix are indexed by the first component, and columns by the second.

3. Theoretical results

3.1 Transforming tensors

In this section we state some technical results related to marginalizations and slices of tensors that arise from stochastic parameters of the general Markov model on a tree T. For a complete proof of these results see [10].

Lemma 3.1. *Let P be a 3-tensor in the image of parameters for the general Markov model, $P = \varphi(\pi, \{M_1, M_2, M_3\})$, where T is a trivalent 3-leaf tree. Then, the three possible marginalizations of P are given by*

$$P_{..+} = M_1^t diag(\pi) M_2, \quad P_{.+.} = M_1^t diag(\pi) M_3, \quad P_{+..} = M_2^t diag(\pi) M_3. \tag{4}$$

And the slices of P are

$$P_{..i} = M_1^T diag(M_3 \mathbf{e_i}) diag(\pi) M_2, \quad P_{.i.} = M_1^T diag(M_2 \mathbf{e_i}) diag(\pi) M_3, \quad P_{i..} = M_2^T diag(M_1 \mathbf{e_i}) diag(\pi) M_3. \tag{5}$$

Corollary 3.2. *Let P be a tensor arising from parameters of the general Markov model on T with tree topology $T_{12|34}$, $P = \varphi_{T_{12|34}}(\pi; M_1, M_2, M_3, M_4, M_5)$ (see the left hand side of Figure 2). Then the double marginalizations $P_{+..+}$, $P_{+.+.}$, $P_{.+.+}$ and $P_{.++.}$ can be computed in terms of the transition matrices as follows:*

$$P_{+..+} = M_2^T diag(\pi) M_5 M_3, \quad P_{+.+.} = M_2^T diag(\pi) M_5 M_4,$$
$$P_{.+.+} = M_1^T diag(\pi) M_5 M_3, \quad P_{.++.} = M_1^T diag(\pi) M_5 M_4.$$

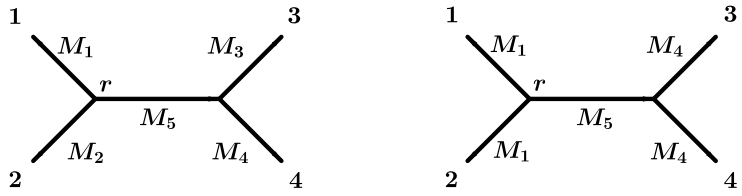

Figure 2: *Left:* Rooted 4-leaf tree $T_{12|34}$ with transition matrices $\{M_1, M_2, M_3, M_4, M_5\}$. *Right:* Rooted 4-leaf tree $T_{12|34}$ with transition matrices $\{M_1, M_1, M_4, M_4, M_5\}$.

The following lemma describes how, given a tensor in the image of φ_T for a 4-leaf tree T, we can produce a new tensor still in $\mathrm{Im}\varphi_T$. This is done by multiplying the original tensor with a matrix (in the sense of (3)), which has the effect of changing the transition matrix of an exterior edge of the tree.

Lemma 3.3. *Let P be a 4-tensor for the general Markov model, $P = \varphi_T(\pi; M_1, \dots, M_5)$. If M_i is non singular for some $i = 1, 2, 3, 4$, then the tensor $\bar{P} = P *_i (M_i^{-1} M)$ is the image of the same parameters as P except for M_i which has been replaced by M.*

3.2 Stochasticity conditions

In this section we will discuss some theoretical results that will allow us to provide some conditions to ensure that a tensor of a joint distribution comes from stochastic parameters.

Definition 3.4. A set $\{\pi, \{M_e\}_{e \in E(T)}\}$ of stochastic parameters for the general Markov model on a tree T with root r is called *nonsingular* if

(i) at every node j of T the distribution of the random variable X_j has no zero entry;

(i) the matrix M_e of every edge e is nonsingular.

Remark 3.5. For stochastic parameters and assuming (ii), condition (i) in the previous definition is equivalent to requiring that the root distribution π_r has no zero entry.

The following result has been proved in [5]. As we do not use it specifically, we do not include the proof here.

Theorem 3.6 (Allman–Rhodes–Taylor, [5])**.** *Let P be a (either real or complex) 3-tensor. Then, P arises from nonsingular parameters for the general Markov model with κ parameters on the 3-leaf tree if and only if the following conditions hold:*

(i) $f_i(P; x) \neq 0$ for an arbitrary vector x and some $i = 1, 2, 3$, where $f_i(P; x) = \det H_x((\det(P *_i x)))$ and H_x denotes the Hessian operator;

(ii) $det(P *_i 1) \neq 0$ for $i = 1, 2, 3$.

We want to find a similar characterization of P for stochastic parameters. That is, we want to find some conditions allowing us to distinguish when a tensor P is the image of positive real parameters.

Theorem 3.7. *Let $P = \varphi_T(\pi, \{M_1, M_2, M_3\})$ be a 3-tensor with π, $\{M_i\}_i$ having real entries. Then,*

(1) P is the image of nonsingular stochastic parameters for the general Markov model on the 3-leaf tree if and only if its components are nonnegative, they sum up to 1, conditions (i) and (ii) from Theorem 3.6 are satisfied, and

(iii) the matrix

$$det(P_{..+})P_{+..}^T adj(P_{..+})P_{.+.} \tag{6}$$

is positive definite, and the following matrices are positive semidefinite for $i = 1, \dots, \kappa$

$$det(P_{..+})P_{i..}^T adj(P_{..+})P_{.+.}, \quad det(P_{..+})P_{+..}^T adj(P_{..+})P_{.i.}, \quad det(P_{+..})P_{.+.} adj(P_{+..})P_{..i}^T. \tag{7}$$

(2) P is the image of nonsingular real positive parameters if and only if its components are positive, they sum up to one, conditions (i) and (ii) are satisfied, and

(iii') all matrices in (6) and (7) are positive definite.

In both cases, the nonsingular parameters are unique up to label swapping.

Proof. The proof of this theorem is essentially the same as in [5], but for real parameters. Let P be an arbitrary nonnegative 3-tensor whose components sum up to 1. Assuming (i) and (ii) and using Theorem 3.6, P is the image of nonsingular parameters. We want to see that condition (iii) is equivalent to these parameters being nonnegative. To this aim, we are going to analyze what is the meaning of expressions (6) and (7).

Let $\bar{P} = P_{+..}P_{..+}^{-1}P_{.+.}$, using expressions proved in Lemma 3.1 we compute

$$\begin{aligned} \bar{P} = P_{+..}^T P_{..+}^{-1} P_{.+.} &= (M_2^T diag(\pi) M_3)^T (M_1^T diag(\pi) M_2)^{-1} (M_1^T diag(\pi) M_3) \\ &= M_3^T diag(\pi) M_3. \end{aligned} \tag{8}$$

This is a well defined symmetric matrix since $P_{..+}$ is nonsingular. Since M_3 is real, $\bar{P}$ is a positive definite matrix if and only if

$$x^T \bar{P} x = x^T M_3^T \mathrm{diag}(\pi) M_3 x = (M_3 x)^T \mathrm{diag}(\pi)(M_3 x) > 0, \quad \forall x \neq 0.$$

Since M_3 is nonsingular, it can be understood as a change of basis and hence $\bar{P}$ is positive semidefinite if and only if the entries of $\mathrm{diag}(\pi)$ are all positive. We clear denominators and obtain an algebraic expression multiplying this matrix by the square of the appropriate nonzero determinant. It follows that (6) is positive definite if and only if π is positive.

Using the expressions in Lemma 3.1, we have

$$\begin{aligned} P_{i..}^T P_{..+}^{-1} P_{.+.} &= (M_2^T diag(M_1 \mathbf{e_i}) M_3)^T (M_1^T diag(\pi) M_2)^{-1} (M_1 diag(\pi) M_3) = \\ &= M_3^T diag(\pi) diag(M_1 \mathbf{e_i}) M_3. \end{aligned}$$

This matrix is also symmetric, and it is positive semidefinite if and only if the entries of $diag(\pi)diag(M_1\mathbf{e_i})$ are nonnegative. Since π is a positive vector, we need the i-th column of M_1 being nonnegative. Using the

matrices $P_{+..}^T P_{..+}^{-1} P_{.i.}$ and $P_{..+}^T P_{+..}^{-1} P_{..i}$ we can also impose the conditions of the i-th column of M_2 and M_3 being nonnegative. This proves (1).

If the matrices of (6) and (7) are positive definite, we can repeat this proof but requiring positiveness of the parameters. This proves (2).

In order to clear denominators and obtain an algebraic expression, we multiply all these matrices by the square of the appropriate nonzero determinant which does not change the sign and gives us expressions (6) and (7). □

Counterexample 3.8. In paper [5], Theorem 3.7 is announced for general tensors P, that is, for $P = \varphi_T(\pi, \{M_1, M_2, M_3\})$ where π, M_1, M_2 and M_3 are complex. But we provide here a counterexample to show that if M_3 is not real, $diag(\pi)$ being positive does not imply $\bar{P} = M^T diag(\pi) M$ being positive definite; see (8). For $\kappa = 2$ let us consider the matrices

$$D = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad M = \frac{1}{4}\begin{pmatrix} 2+i & 2-i \\ 2-i & 2+i \end{pmatrix}.$$

However, the matrix $M^T D M = \frac{1}{16}\begin{pmatrix} 3 & 5 \\ 5 & 3 \end{pmatrix}$ is not positive definite.

Moreover, the reverse implication is not true either. For instance, for the positive definite matrix

$$\bar{P} = M^T D M = \begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix},$$

we have the following decomposition, where D is not positive: $D = \begin{pmatrix} -1 & 0 \\ 0 & 4 \end{pmatrix}$, $M = \begin{pmatrix} 2i & -2i \\ 1 & 1 \end{pmatrix}$.

Due to this counterexample we are forced to restrict the statement of the above theorem to the case of real matrices.

Assuming now that an n-tensor P arises from nonsingular parameters on a tree, we would like to give some semialgebraic conditions that are satisfied if and only if P comes from stochastic parameters. If we consider marginalizations of P to three variables and using Theorem 3.7, we can derive conditions that hold when the root distribution and the product of matrices associated to any path from an interior node to a leaf are stochastic. Nevertheless, we need some extra conditions to guarantee matrices of the interior edges being stochastic.

The following result gives us a condition for all parameters of the 12|34 tree being stochastic.

Theorem 3.9 (Allman–Rhodes–Taylor, [5]). *Let P be a 4-tensor. Suppose P arises from nonsingular real parameters for the general Markov model on $T_{12|34}$. If the marginalizations $P_{+...}$ and $P_{...+}$ arise from stochastic parameters and, moreover, the $\kappa^2 \times \kappa^2$ matrix*

$$det(P_{+..+}) det(P_{.+.+}) Flatt_{13|24}\Big(P *_2 \big(adj(P_{+..+}^T) P_{.+.+}^T\big) *_3 \big(adj(P_{.+.+}) P_{.++.}\big)\Big) \tag{9}$$

is positive semidefinite, then P arises from stochastic parameters.

Proof. The root r is placed at the interior node near leaves 1 and 2, as we can see in the tree presented on the left of Figure 2. Let M_i, $i = 1, 2, 3, 4$, be the complex matrix associated to the edges leading to

leaves, M_5 the matrix on the internal edge, and π the root distribution. The rows of these matrices sum up to 1. We define the four matrices

$$\begin{array}{ll} N_{32} = P^T_{+..+} = M_3^T M_5^T \, diag(\pi) M_2, & N_{31} = P^T_{.+.+} = M_3^T M_5^T \, diag(\pi) M_1, \\ N_{14} = P_{.++.} = M_1^T \, diag(\pi) M_5 M_2, & N_{13} = P_{.+.+} = M_1^T \, diag(\pi) M_5 M_3. \end{array} \tag{10}$$

We define now a tensor $\bar{P}$ arising from the same parameters as P except that M_2 has been replaced by M_1 (see Lemma 3.3) and, similarly, a tensor $\tilde{P}$ arising from the same parameters as $\bar{P}$ but with M_4 instead of M_3:

$$\bar{P} = P *_2 N_{32}^{-1} N_{31} = P *_2 M_2^{-1} M_1, \quad \tilde{P} = \bar{P} *_3 N_{13}^{-1} N_{14} = \bar{P} *_3 M_3^{-1} M_4. \tag{11}$$

We can express

$$Flat_{13|24}(P) = (M_1 \otimes M_3)^T D (M_2 \otimes M_4), \tag{12}$$

where D is the diagonal matrix containing the κ^2 entries of $diag(\pi)M_5$; see [10] for further details. Since $\tilde{P}$ arises from the same parameters that P except that M_2 has been replaced by M_1 and M_3 by M_4, we can write $Flat_{13|24}(\tilde{P}) = (M_1 \otimes M_4)^T D (M_1 \otimes M_4)$.

Since the 3-marginalization arises from stochastic parameters, M_1 and M_4 are nonsingular and the components of π are positive. Thus, $M_1 \otimes M_4$ is also nonsingular. All principal minors of $Flat_{13|24}(\tilde{P})$ are nonnegative if and only if $Flat_{13|24}(\tilde{P})$ is positive semidefinite. Then we have to require the entries of D to be nonnegative and so, since π has positive components, we can ensure that M_5 has nonnegative entries. By multiplying $Flat_{13|24}(\tilde{P})$ by the square of the appropriate nonzero determinant, we clear denominators and obtain the algebraic expressions stated in the theorem. □

Remark 3.10. The theoretical results proved in this section complement the algebraic description of the model (given by topology invariants) with a semialgebraic description of the points with stochastic sense. In other words, as well as finding polynomials vanishing on the image of the parametrization map, we have found polynomial inequalities sufficing to characterize the stochastic image.

The conditions of matrices being positive definite/semidefinite can be expressed as semialgebraic conditions using Sylvester's criterion, which claims that a real symmetric matrix is positive definite (resp., positive semidefinite) if and only its *leading* principal minors are positive (resp., nonnegative).

On the other hand, the replacements of inverses in (11) by adjoint matrices in (9) is not only done in order to have semialgebraic conditions, but also to avoid dealing with the inverse of ill conditioned matrices.

Let P be the tensor used in Theorem 3.9 and $\tilde{P}$ the one constructed in (11). Since $\tilde{P}$ arises from the same parameters that P except that M_2 has been replaced by M_1 and M_3 by M_4, it is the joint distribution of the tree presented on the right hand side of Figure 2. Observing the symmetry of the exterior transition matrices we can state the following result.

Theorem 3.11. *Let P be a 4-tensor whose components sum up to 1. Suppose that*

$$P = \varphi_T(\pi, M_1, M_2, M_3, M_4, M_5),$$

with $T = T_{12|34}$, and let $\tilde{P}$ be constructed as in (11). Then,

$$Flat_{13|24}(\tilde{P}) = Flat_{14|23}(\tilde{P}) \quad \text{and} \quad Flatt_{12|34}(\tilde{P}) \neq Flatt_{13|24}(\tilde{P}). \tag{13}$$

In particular, the equality of matrices

$$\begin{aligned} & det(P_{+..+})det(P_{.+.+})Flatt_{13|24}\Big(P *_2 \left(adj(P_{+..+}^T)P_{.+.+}^T\right) *_3 \left(adj(P_{.+.+})P_{.++.}\right)\Big) = \\ = \; & det(P_{+..+})det(P_{.+.+})Flatt_{14|23}\Big(P *_2 \left(adj(P_{+..+}^T)P_{.+.+}^T\right) *_3 \left(adj(P_{.+.+})P_{.++.}\right)\Big) \end{aligned}$$

gives rise to 256 *topology invariants of degree* 17*.*

Proof. Using (12), and the fact that, in $\tilde{P}$, M_2 has been replaced by M_1, and M_3 by M_4, we have

$$Flat_{13|24}(\tilde{P}) = (M_1 \otimes M_4)^T D(M_1 \otimes M_4) = Flat_{14|23}(\tilde{P}). \tag{14}$$

In contrast, $Flatt_{12|34}(\tilde{P}) = \bar{M}_1^T diag(\pi)\bar{M}_4$, where $\bar{M}_1(x_i,(x_j,x_k)) = M_1(x_i,x_j)M_1(x_i,x_k)$, $\tilde{M}_4(x_i,(x_j,x_k)) = \sum_{l=1}^{\kappa} M_5(x_i,x_l)M_4(x_l,x_j)M_4(x_l,x_k)$, is, in general, not equal to (14).

The expression $Flat_{13|24}(\tilde{P}) = Flat_{14|23}(\tilde{P})$ provides 16×16 equalities between entries. By (9), these entries are algebraic expressions in terms of components of P. Moreover, because of (13), these equalities are not satisfied by any distribution arising from a tree and then they are topology invariants.

Finally, regarding (9), we infer the degree of these expressions in the components of P:

(i) the two determinants have degree 4 each, which makes degree 8;

(ii) the components of the tensors $adj(P_{+..+}^T)P_{.+.+}^T$ and $adj(P_{.+.+})P_{.++.}$ have degree 4.

The $*$ operation adds degrees, so we obtain a tensor of degree $1+4+4=9$ before applying $Flat_{13|24}(\cdot)$. Altogether gives a tensor with components of degree $8+9=17$. □

4. Conclusions

In this paper, we have seen that the conditions of stochasticity on the parameters from Theorem 3.9 are enough to ensure that the 4-tensor arising from real nonsingular parameters under the general Markov model comes from stochastic parameters. From these conditions we have been able to find new topology invariants. So, we can extract the following conclusions:

(i) we have disentangled the theoretical results of stochastic conditions of the parameters and we have provided a counterexample to an error in a proof of [5] as well;

(ii) using the ideas from the proof of Theorem 3.9 we have provided 256 topology invariants of degree 17.

However, there is still further research to do:

(i) check whether the new topology invariants we found are sufficient to describe the phylogenetic algebraic variety;

(ii) check if these conditions can be used with real data, in order to give new information that can be used in some phylogenetic reconstruction method.

References

[1] E.S. Allman and J.A. Rhodes, "Phylogenetic invariants for the general Markov model of sequence mutation", *Math. Biosci.* **186**(2) (2003), 113–144.

[2] E.S. Allman and J.A. Rhodes, "Mathematical models in biology, an introduction", Cambridge University Press (2004). ISBN 0-521-52586-1.

[3] E.S. Allman and J.A. Rhodes, "The mathematics of phylogenetics", University of Alaska Fairbanks (2005).

[4] E.S. Allman and J.A. Rhodes, "Phylogenetic invariants", in *Reconstructing evolution*, Oxford Univ. Press, Oxford (2007), 108–146.

[5] E.S. Allman, J.A. Rhodes, and A. Taylor, "A semialgebraic description of the general Markov model on phylogenetic trees", *Preprint* (2012), `http://adsabs.harvard.edu/abs/2012arXiv1212.1200A`.

[6] M. Casanellas, "Algebraic tools for evolutionary biology", *La Gaceta de la RSME* **15** (2012), 521–536.

[7] M. Casanellas and J. Fernández-Sánchez, "Reconstrucción filogenética usando geometría algebraica", *Arbor. Ciencia, pensamiento, cultura* **96** (2010), 207–229.

[8] M. Casanellas and J. Fernández-Sánchez, "Relevant phylogenetic invariants of evolutionary models", *Journal de Mathématiques Pures et Appliquées* **96** (2011), 207–229.

[9] N. Eriksson, "Tree construction using singular value decomposition", in L. Pachter and B. Sturmfels, editors, "Algebraic Statistics for computational biology" (chapter 19), Cambridge University Press (2005), 347–358.

[10] M. Garrote, "Multilinear algebra for phylogenetic reconstruction", Master's thesis, Universitat Politècnica de Catalunya (2015).