BACHELOR'S DEGREE THESIS

Degree in Mathematics

Title: Phylogenetics and rank of matrices

Author: Marina Garrote López

Advisor: Marta Casanellas

Department: Matemàtica Aplicada I

Academic year: 2013 - 2014

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Bachelor's degree thesis

Phylogenetics and rank of matrices

Marina Garrote López

Advisor: Marta Casanellas Rius

Departament de Matemàtica Aplicada I

To all the people who have placed their trust in me and day after day have given me the courage I needed to do it.

Contents

Introduction

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

The aim of this work is to study the relationship between phylogenetics and these algebraic techniques. But what does phylogenetics study? Phylogenetics is the study of the evolutionary relationships of a group of species, and it is usually inferred from the DNA sequences of a set of living species.

Thanks to the model developed by Darwin based on natural selection, we can construct *phylogenetics trees* that relate a set of contemporary species. Thus phylogenetics tries to reconstruct these trees. In order to do 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 our model, known as *phylogenetic invariants*. Mathematicians have recently begun to be interested in the study of these polynomials and the study of the geometry of the algebraic varieties that arise in this setting.

The main goal of this work is to understand the relationship between phylogenetics and these algebraic techniques and to prove that using the rank of certain matrices we can find phylogenetic invariants that are useful for tree reconstruction. To this end we have adapted a proof from a result of Allman and

Rhodes and have extended it in order to prove that these rank conditions give indeed phylogenetic invariants. We also relate the results to the framework of multininear algebra by understanding joint distributions as tensors.

This work is divided into two parts. In the first part we explain concepts that are already known. We will explain what are phylogenetic trees from the mathematical standpoint and we will present several models for these trees. Once we have studied the models we will explain what phylogenetic invariants are and how these can be found, and see how they can be used to reconstruct the structures of trees. All these concepts are explained in Chapter 1 and in the first section of the second chapter.

In the second part of this work we develop our contribution in understanding the results and proofs of Allman and Rhodes. In this part, presented in the second section of Chapter 2 we will formulate and proof two important results that provide phylogenetic invariants for trees generated by four species. In order to complement these results we have added a computer program that with specific cases corroborates our results (see Annex A).

Chapter 1

Evolutionary models

1.1 Phylogenetic trees

1.1.1 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 this theory with phylogenetic trees (see Fig.1.1). The nodes of this tree 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 will the common ancestor of all species.

Genetic information of each individual is encoded in the DNA of the nucleus of its cells. DNA molecules are composed of simpler units called nucleotides and consist of two anti-parallel strands of nucleotides coiled around each other to form a double helix. Each nucleotide is composed of a phosphate, a sugar and a basis. According to the bases, nucleotides are called adenine (A), cytosine (C), guanine (G) and thymine (T). A base-pair is one of the pairs $\mathtt{A}-\mathtt{T}$ or $\mathtt{C}-\mathtt{G}$. The nucleotides on a base-pair are complementary in the sense that in the double helix adenine connects with the thymine and the guanine with

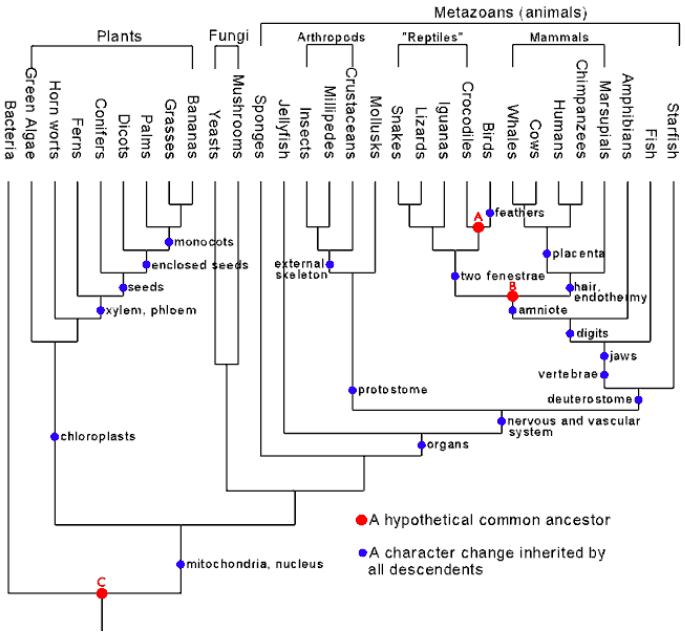

Figure 1.1: A phylogenetic tree

cytosine. According to this symmetry, we store a DNA molecule as an ordered sequence of `A`, `C`, `G` and `T` (see Fig. 1.2).

The heredity information in a genome is thought to be contained in the genes. But the DNA sequences of a same gene may not be the same for different species. They contain similar parts but they can also contain some other parts that we can not compare. For that reason the first problem is identifying which part of the DNA sequences of different species we can compare. 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 sequences. We can represent the alignment with a table whose rows are the species DNA sequences and whose columns correspond to nucleotides that have evolved from the same nucleotide of the common ancestor of all the species (see Table 1.1). Alignments are used in many contexts, in phylogenetics among them, to see relationships between some species and to reconstruct the phylogenetic tree that relates them. Changes in DNA sequences of different

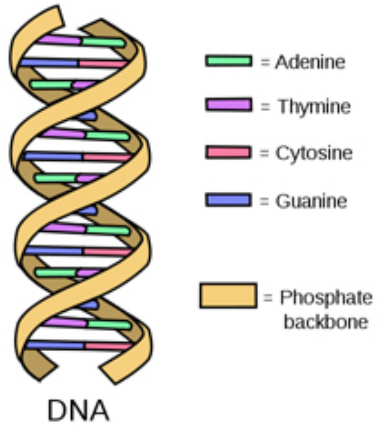

Figure 1.2: DNA molecule

species are given by substitutions, insertions or deletions. In the two latter cases, a nucleotide is inserted or deleted from a given position as compared with the other sequence. In most commonly used evolutionary models, insertions and deletions are not considered and incorporating them would highly increase the complexity of the model. So in this work we will assume that mutations in different alignments are just substitutions. Therefore the alignments we will deal with have the same length and contain no gaps.

Gorilla Gorilla	AACTTCGAGGCTTACCGCTG
Homo Sapiens	AACGTCTATGCTCACCGATG
Pan Troglodytes	AAGGTCGATGCTCACCGATG

Table 1.1: A multiple sequence alignment of DNA sequences of *Homo Sapiens* (Human), *Pan Troglodytes* (Chimpanzee) and *Gorilla Gorilla* (Gorilla).

1.1.2 Phylogenetic trees as graphs

In this section we introduce the concepts that allow us to deal with phylogenetic trees. We follow the approach in [AR04], [AR05] and [Cas12].

Definition 1.1.1. A *tree* $\mathcal{T}$ is a connected acyclic graph. The *degree* of a vertex in a tree $\mathcal{T}$ is the number of edges incident to it. A *leaf* is a vertex of degree 1 while an *internal vertex* is a vertex of degree at least 2. The set of leaves is usually denoted by $L(\mathcal{T})$ and leaves are labelled from 1 to n. $Int(\mathcal{T})$ denote the set of interior nodes and $E(\mathcal{T})$ the set of edges of the tree. If all nodes in $Int(\mathcal{T})$ have degree 3, then $\mathcal{T}$ is called a *trivalent tree*.

Definition 1.1.2. A tree is called a *rooted tree* if one vertex has been labelled as "root", and the edges are oriented away from the root.

Definition 1.1.3. Let X denote a finite set of labels. Then a *phylogenetic tree* is a pair $(\mathcal{T}, \phi)$ where $\mathcal{T}$ is a tree and $\phi : X \to L(\mathcal{T})$ is a one-to-one correspondence.

In a phylogenetic tree the set X represents a set of living species and the tree $\mathcal{T}$ shows the ancestral relationships among them. If the phylogenetic tree is rooted, then the root represents the common ancestor to the set of species X.

Example 1.1.4. Figure 1.3 is an example of a phylogenetic tree, where $X = \{$*Homo sapiens, Gorilla Gorilla, Pan Troglodytes, Macaca Mulatta*$\}$ is the set of labels.

The correspondence between the set of leaves and the set of labels is:

$$\begin{array}{ccc} L(\mathcal{T}) & \overset{\phi}{\longleftrightarrow} & X \\ 1 & \overset{\phi}{\longleftrightarrow} & Homo\ Sapiens \\ 2 & \overset{\phi}{\longleftrightarrow} & Gorilla\ Gorilla \\ 3 & \overset{\phi}{\longleftrightarrow} & Pan\ Troglodytes \\ 4 & \overset{\phi}{\longleftrightarrow} & MacacaMulatta \end{array}$$

For our purposes, usually the set X will coincide with the set $\{1, 2, \ldots, n\}$. A phylogenetic rooted tree has an induced orientation on its edges. The length of the edges, called *branch length* represents the evolutionary distance. This can be represented for example by the number of nucleotide changes that have

occurred along the evolutionary process represented by that branch between two species.

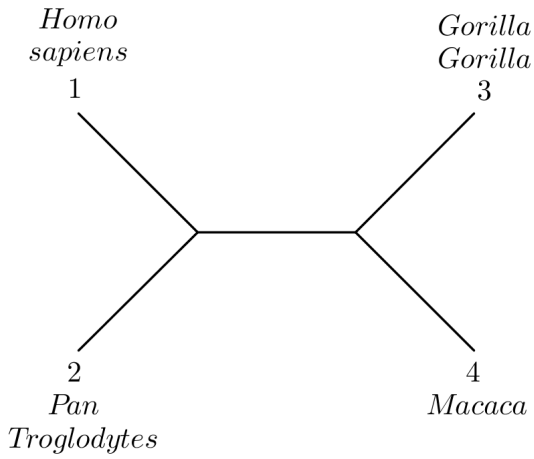

Figure 1.3: 4-leaf phylogenetic tree on the set of species *Homo Sapiens* (human), *Pan Troglodytes* (chimpanzee), *Gorilla Gorilla* (gorilla), and *Macaca Mulatta* (macaque).

Definition 1.1.5. The *tree topology* of a phylogenetic tree is the topology as a labelled graph.

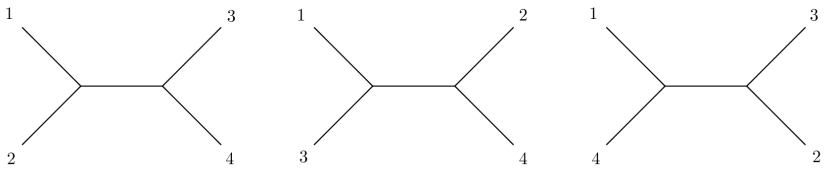

Figure 1.4: The 3 possible topologies for T_4

That is, two phylogenetic trees with the same set of labels X at the leaves, $\mathcal{T}_1$ and $\mathcal{T}_2$, have the same topology if there is a one-to-one correspondence φ

between their vertices that respects adjacency and their leaf labelling. If $\mathcal{T}_1$ and $\mathcal{T}_2$ are rooted trees and $r_1,\ r_2$ are their roots respectively then we need to impose $\varphi(r_1) = (r_2)$.

We will denote by T_n the set of all possible possible unrooted trivalent tree topologies for n-leaf trees. For example Figure 1.4 shows the 3 possible topologies for T_4.

Example 1.1.6. Tree represented in Figure 1.5 has the same topology as the tree represented in Figure 1.6 if it is considered as an unrooted tree.

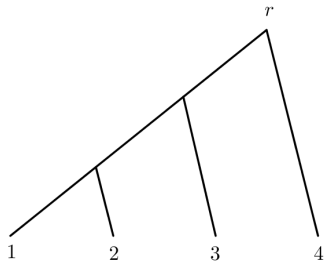

Figure 1.5: 4-leaf rooted tree

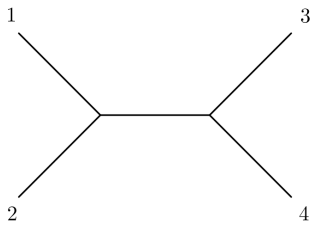

Figure 1.6: 4-leaf unrooted tree

The first goal in phylogenetics is, given an alignment of n DNA sequences of n different species, infer which of the T_n topologies explains best the evolution of this set of species. Another goal in phylogenetics is to infer the branch lengths on this tree (evolutionary distance), but we will not deal with this problem in this work.

1.2 Evolutionary models

One usually models evolution 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.

Let $\mathcal{T}$ be a rooted phylogenetic tree with n leaves. We assume that the nucleotides in the DNA sequence are independent and identically distributed (iid). That is, the states at each position in the sequence evolve independently of the other nucleotides and according to the same evolutionary process. To this end we associate a discrete random variable X_i to each node i of $\mathcal{T}$ with κ possible states. Let $\mathcal{K}$ be the set of the κ states. Usually $\kappa = 4$ and the states represent the four nucleotides in DNA, *Adenine*, *Cytosine*, *Guanine* and *Thymine* and then $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Random variables at the leaves are observed because the DNA sequences of the contemporary species are observations of the variables, while the random variables at the interior nodes are hidden. If $\mathcal{T}$ has leaves $1, 2, \ldots, n$, then $X = (X_1, X_2, \ldots, X_n)$ is the vector of joint distribution at the leaves and each column of the alignment is an observation of this vector of random variables.

Let $\pi = (\pi_1, \ldots, \pi_\kappa)$ be the distribution of X_r at the root r (all entries are nonnegative and $\sum_i \pi_i = 1$). If $\kappa = 4$ and $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ then we interpret these entries as giving the probabilities that an arbitrary site in the DNA sequence at the root is occupied by the corresponding base, or, equivalently, as the frequencies with which we would expect to observe these bases in a sequence at the root. Now for each edge e we associate a $\kappa \times \kappa$ matrix M_e. We will call this matrix *substitution* or *transition* matrix. Its entries are the conditional probabilities $P(x|y, e)$ of a state y at the parent node of e being substituted by a state x at its child, during the evolutionary process along branch e. Thus the rows of M_e sum to 1 and M_e is a *Markov matrix* (or a row *stochastic matrix*). If $\kappa = 4$ and $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ then the (i, j)-entry of M_e stands for the conditional probability that if nucleotide i occurs at one site in the parent vertex on the edge e, then nucleotide j occurs at the descendent vertex at the same site.

Therefore a substitution matrix for $\kappa = 4$ and $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ is

$$M_e = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} P(\mathtt{A}|\mathtt{A}, e) & P(\mathtt{C}|\mathtt{A}, e) & P(\mathtt{G}|\mathtt{A}, e) & P(\mathtt{T}|\mathtt{A}, e) \\ P(\mathtt{A}|\mathtt{C}, e) & P(\mathtt{C}|\mathtt{C}, e) & P(\mathtt{G}|\mathtt{C}, e) & P(\mathtt{T}|\mathtt{C}, e) \\ P(\mathtt{A}|\mathtt{G}, e) & P(\mathtt{C}|\mathtt{G}, e) & P(\mathtt{G}|\mathtt{G}, e) & P(\mathtt{T}|\mathtt{G}, e) \\ P(\mathtt{A}|\mathtt{T}, e) & P(\mathtt{C}|\mathtt{T}, e) & P(\mathtt{G}|\mathtt{T}, e) & P(\mathtt{T}|\mathtt{T}, e) \end{pmatrix} \end{array}.$$

The probabilistic model we have described is a Markov process in the following sense.

Definition 1.2.1. A Markov process is a random phenomenon, such that it complies the Markov property that says that "the process has no memory", which means that the probability distribution of the future value of a variable depends on its present value, but is separated from the history of the variable.

In other words, in a Markov process the probability that a particular state change occurs given the system is in state i is the same as the probability of the same change, given any entire earlier history of states ending in state i.

The model we have explained above, of molecular evolution occurring through random nucleotides substitutions satisfies the Markov assumption. Since the probabilities of the various possible state changes on any given edge depend only on the state at the ancestral node on that edge. Besides, we only have observations of the random variables at the leaves and we do not have observations for the variables at the interior nodes, so ours is a *hidden* Markov process.

Example 1.2.2. In Figure 1.7 let X_1 the random variable associated to *Gorilla*, X_2 displays nucleotides in *human* and X_3 in *Chimpanzee*. Therefore the random vector $X = (X_1, X_2, X_3)$ represents the observed random variables associated with the Gorilla, the Human and the Chimpanzee respectively. For example the observation of the vector X associated to the first column of the alignment in Table 1.1.1 is $(\mathtt{A},\mathtt{A},\mathtt{C})$. We associate a *transition* matrix M_e to each edge of the rooted phylogenetic tree that connects the three species.

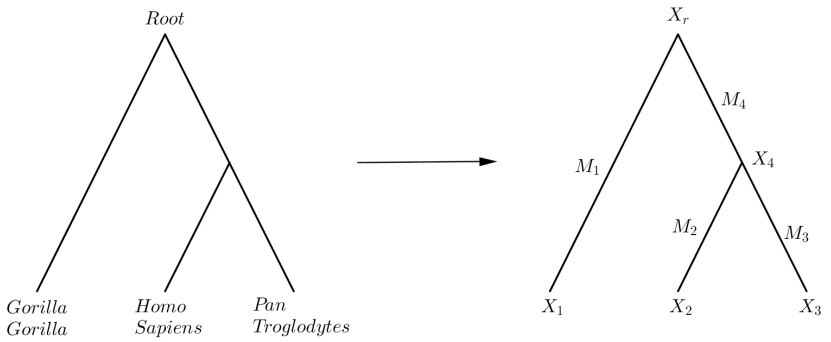

Figure 1.7: *Left:* Phylogenetic tree of *Gorilla Gorilla, Pan troglodytes* and *Homo Sapiens*. *Right:* Statistical model on a rooted phylogenetic 3-leaved tree.

According to the shape of the transition matrices one has different *models*. If we do not impose any restriction, then we have the model called *general Markov model GMM* where the substitution matrices M_e are of the form

$$M_e = \begin{pmatrix} a_e & b_e & c_e & d_e \\ e_e & f_e & g_e & h_e \\ j_e & k_e & l_e & m_e \\ n_e & o_e & p_e & q_e \end{pmatrix}, \ where \ \begin{cases} a_e + b_e + c_e + d_e = 1 \\ e_e + f_e + g_e + h_e = 1 \\ j_e + k_e + l_e + m_e = 1 \\ n_e + o_e + p_e + q_e = 1 \end{cases}$$

Now we present some other models, which are more restrictive than the GMM.

Definition 1.2.3. *Jukes Cantor model.*
This is the most restricted model since it adds several additional assumptions to the general Markov model. At the same time is really simple. First of all it assumes that all bases occurring with equal probability in the ancestral

sequence. Therefore the root distribution vector is

$$\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right).$$

And now it assumes the substitution matrices are of the form

$$M_e = \begin{pmatrix} a_e & b_e & b_e & b_e \\ b_e & a_e & b_e & b_e \\ b_e & b_e & a_e & b_e \\ b_e & b_e & b_e & a_e \end{pmatrix},$$

and since the rows sum to 1, $b_e = (1 - a_e)/3$.

Definition 1.2.4. *Strand symmetric model.*
Another model that has a particular interest is the Strand symmetric model that reflects the double strand symmetry of DNA molecules. As we have explained, in the DNA molecule nucleotides are linked in pairs $\mathtt{A} - \mathtt{T}$ and $\mathtt{C} - \mathtt{G}$, so this model contemplates this and assumes the following restrictions $j_e = h_e$, $k_e = g_e$, $l_e = f_e$, $m_e = e_e$, $n_e = d_e$, $o_e = c_e$, $p_e = b_e$, $q_e = a_e$ (see the GMM matrix in the previous page), $\pi_A = \pi_T$ and $\pi_C = \pi_G$. Therefore the matrices are,

$$M_e = \begin{pmatrix} a_e & b_e & c_e & d_e \\ e_e & f_e & g_e & h_e \\ h_e & g_e & f_e & e_e \\ d_e & c_e & b_e & a_e \end{pmatrix},$$

with sum of rows equal to 1.

Definition 1.2.5. *Kimura models.*
Kimura 3-parameter is a model introduced by Kimura in 1981. This model assumes that the base frequencies at the root are equal. It is more flexible than Jukes Cantor model since in this case it has three free parameters. The

transition matrices are

$$\begin{pmatrix} a_e & b_e & c_e & d_e \\ b_e & a_e & d_e & c_e \\ c_e & d_e & a_e & b_e \\ d_e & c_e & b_e & a_e \end{pmatrix},$$

where $a_e = 1 - b_e - c_e - d_e$ and the root distribution is assumed to be uniform.

A more restricted model is the Kimura 2-parameter model, which assumes moreover that $b_e = d_e$.

1.3 Evolutionary models and polynomial maps

We fix now an evolutionary model $\mathcal{M}$ on a tree $\mathcal{T}$ of n leaves rooted at a node r. We call $\mathcal{K}$ the set of states of the random variables at the nodes of the tree and let κ be the cardinal of $\mathcal{K}$. In what follows we can describe how to compute the joint probability of observing states $x_1, x_2, \ldots, x_n$ at the leaves according to the Markov process we have described.

We will denote by $p_{x_1,\ldots,x_n}$ the joint distribution at the leaves of a rooted phylogenetic tree $\mathcal{T}$, that is, $p_{x_1,\ldots,x_n}$ is the probability that leaves $1, \ldots, n$ take the states $x_1, \ldots, x_n$:

$$p_{x_1,\ldots,x_n} = Prob(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

We define P as the vector whose entries are the joint probabilities $p_{x_1\ldots x_n}$,

$$P = (p_{x_1\ldots x_n})_{x_1\ldots x_n \in \mathcal{K}}.$$

Then as the evolutionary processes are independent and just depend on a common node we can express $p_{x_1,\ldots,x_n}$ in terms of the entries of the substitution matrices. Then we can compute $p_{x_1,\ldots,x_n}$ in the following way.

$$p_{x_1,\ldots,x_n} = \sum_{x_r, (x_v)_{v \in Int(\mathcal{T})}} \prod_{e \in E(\mathcal{T})} M_e(x_{a(e)}, x_{d(e)}), \tag{1.1}$$

where $x_r \in \mathcal{K}$ is a state of the root, $x_{a(e)} \in \mathcal{K}$ is a state of the node ancestor of the edge e, and $x_{d(e)} \in \mathcal{K}$ is the state of the descendent node of the edge e and if e is a terminal edge to a leaf i then $x_{d(e)} = x_i$.

Example 1.3.1. Let $\mathcal{T}$ be the 5-leaf tree of Figure 1.8. Suppose every random variable has $\mathcal{K}$ as the set of states. Considering the Figure 1.8 with M_i are the substitution matrices for each e_i.

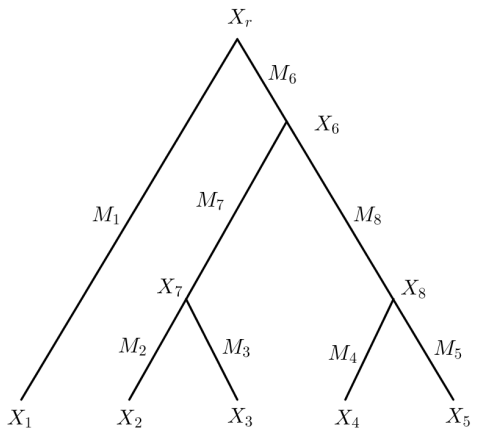

Figure 1.8: Statistical model on a rooted phylogenetic 5-leaved tree.

Then the joint distribution $p_{x_1,x_2,x_3,x_4,x_5} = Prob(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5)$ for this tree in terms of the transition matrices is

$$p_{x_1,x_2,x_3,x_4,x_5} = \sum_{x_r \in \mathcal{K}} \sum_{x_6 \in \mathcal{K}} \sum_{x_7 \in \mathcal{K}} \sum_{x_8 \in \mathcal{K}} \pi_{x_r} M_1(x_r, x_1) M_2(x_7, x_2) M_3(x_7, x_3) \times$$

$$\times M_4(x_8, x_4) M_5(x_8, x_5) M_6(x_r, x_6) M_7(x_6, x_7) M_8(x_6, x_8).$$

Example 1.3.2. Consider the Example 1.2.2. Then $p_{xyz} = Prob(X_1 = x, X_2 = y, X_3 = z)$ is the probability of observing nucleotides x, y, z at the leaves *Gorilla Gorilla*, *Homo sapiens* and *Pan troglodytes* respectively. In

terms of the transition matrices of the model $\mathcal{M}$, we have

$$p_{xyz} = \sum_{x_r, x_4 \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_{x_r} M_1(x_r, x) M_4(x_r, x_4) M_2(x_4, y) M_3(x_4, x), \tag{1.2}$$

where $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})$ is the distribution of nucleotides at the root. Consider that the model $\mathcal{M}$ is the Jukes Cantor model, then

$$p_{\mathtt{AAA}} = \frac{1}{4}(a_1a_4a_2a_3 + 3b_1b_4a_2a_3 + 3b_1a_4a_2a_3 + 3a_1b_4a_2a_3 + 6b_1b_4b_2b_3),$$

$a_i + 3b_i = 1$ for $i = 1, 2, 3, 4$.

On the other hand the probability that the tree $\mathcal{T}$ of figure 1.7 has produced the alignment in table 1.1 equals

$$(p_{\mathtt{AAA}})^4 * p_{\mathtt{CCG}} * p_{\mathtt{TGG}} * (p_{\mathtt{TTT}})^3 * (p_{\mathtt{CCC}})^4 * p_{\mathtt{GTG}} * p_{\mathtt{GTT}} * (p_{\mathtt{GGG}})^3 * p_{\mathtt{TCC}} * p_{\mathtt{CAA}}.$$

We will study the relations among the joint distributions entries. To this end, we view $P = (p_{x_1,\ldots,x_n})_{x_1,\ldots,x_n}$ as a vector space in $\mathbb{C}^{\kappa^n}$. Let $\mathcal{T}$ be a rooted phylogenetic tree and $\mathcal{M}$ an evolutionary model. Let r be the root of $\mathcal{T}$ and X_i the random variables associated to the n leaves that can take κ different states from the set $\mathcal{K} = \{s_1, \ldots, s_\kappa\}$. Then we can define a map $\varphi_T^{\mathcal{M}} : \mathbb{R}^d \to \mathbb{R}^{4^n}$ that sends the set of d parameters to the set of the κ^n possible joint distribution at the leaves. It may seem strange that we define this map in $\mathbb{R}$ or $\mathbb{C}$ when we were talking about probabilities (so real numbers in [0,1]), but we want to define this application on $\mathbb{C}$ in order to use techniques from algebraic geometry.

More formally this map is

$$\begin{array}{rcl} \varphi_T^{\mathcal{M}} : \mathbb{C}^d & \longrightarrow & \mathbb{C}^{\kappa^n} \\ (\pi, \{M_e\}_{e \in E(\mathcal{T})}) & \mapsto & p = (p_{s_1 s_1 \ldots s_1}, p_{s_1 s_1 \ldots s_2}, p_{s_1 s_1 \ldots s_3}, \ldots, p_{s_\kappa s_\kappa \ldots s_\kappa}), \end{array}$$

where $p_{x_1 \ldots x_n}$ is expressed in terms of the root distribution π and the transition matrices M_e according to the expression 1.1 and d stands for the number of free parameters of the model. Although both parameters and image points stand for probabilities we allow them to belong to $\mathbb{C}$ (and not only in the

simplex) because this is enough to deduce algebraic equations satisfied by the image points.

It can be proved that if we root the tree $\mathcal{T}$ at a different node r' (call this tree $\mathcal{T}'$) then, for any set of parameters $\pi, \{M_e\}_{e\in E(\mathcal{T})}$, there exist parameters $\pi, \{M_e\}_{e\in E(\mathcal{T})}$ such that

$$\phi_T^{\mathcal{M}}(\pi, \{M_e\}_e) = \phi_{T'}^{\mathcal{M}}(\pi', \{M'_e\}_e).$$

This means that the root position cannot be inferred from the joint distribution at the leaves. This phenomenon is usually known as the non-identifiability of the root position.
This is why we usually deal with unrooted trees when addressing the problem of topology reconstruction.

Let T be the phylogenetic tree topology of the Figure 1.7 and consider the Jukes Cantor model. Suppose $\kappa = 4$ and $\mathcal{K} = \{A, C, G, T\}$. Then the map that corresponds to this topology is:

$$\begin{array}{rcl} \varphi_T^{JC} : \mathbb{C}^4 & \longrightarrow & \mathbb{C}^{64} \\ (a_1, a_2, a_3, a_4) & \mapsto & p = (p_{\texttt{AAA}}, p_{\texttt{AAC}}, p_{\texttt{AAG}}, \ldots, p_{\texttt{TTT}}), \end{array}$$

and the joint probabilities at the leaves are written in terms of the parameters using expression 1.1.

1.4 Joint distribution as a tensor

Although it is not essential for the remaining sections, we now introduce a more algebraic way of viewing the joint distribution at the leaves of a phylogenetic tree, which was one of our goals in this project.

Let $\mathcal{T}$ be a rooted phylogenetic tree and $\mathcal{M}$ an evolutionary model. Let r be the root of $\mathcal{T}$ and X_i the random variables associated to the n leaves that can take κ different states from the set $\mathcal{K} = \{s_1, \ldots, s_\kappa\}$.

Let $\mathcal{W} := \mathbb{C}^\kappa$ be a vector space. We identify the canonical basis of $\mathcal{W}$ with the set $\mathcal{K}$. Then the natural basis of $\mathcal{W} \otimes \overset{n)}{\ldots} \otimes \mathcal{W}$ is $\{x_1 \otimes \ldots \otimes x_n\}_{x_1,\ldots,x_n\in\mathcal{K}}$.

For example if $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, then the natural basis of $\mathcal{W} \otimes \mathcal{W} \otimes \mathcal{W}$ is $\{\mathtt{A} \otimes \mathtt{A} \otimes \mathtt{A},\ \mathtt{A} \otimes \mathtt{A} \otimes \mathtt{C},\ \ldots,\ \mathtt{T} \otimes \mathtt{T} \otimes \mathtt{T}\}$

The joint distribution $P = (p_{x_1 \ldots x_n})_{x_1 \ldots x_n \in \mathcal{K}}$ can be thought as a $\kappa \times \overset{n)}{\ldots} \times \kappa$ array and p can be viewed as a tensor in $\mathcal{W} \overset{n)}{\otimes \ldots} \otimes \mathcal{W}$ whose coordinates in the natural basis above are $p = (p_{x_1 \ldots x_n})_{x_1 \ldots x_n \in \mathcal{K}}$.

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

This formulation of the joint distribution arises naturally as we will observe in the following chapter.

Chapter 2

Phylogenetic invariants for tree reconstruction

In this section we will explain that we can see the evolutionary models that we described in the section 1.2 as algebraic varieties, and we will see how algebraic geometry can help us to reconstruct phylogenetic trees. We follow the results presented in [Cas12], [CFS10], [Eri05] and [AR07]. To be able to reconstruct these trees it can be useful know algebraic relationships among coordinates of the joint distribution at the leaves. We will be interested in know the algebraic relations that are satisfied by the joint distribution of a tree topology but not for the others topologies of the same tree. This is what biologists Cavender and Falsestein called in 1987 *phylogenetic invariants* (see Def. 2.1.4) and they indicated this as a potential tool of the reconstruction of the tree topology.

2.1 Phylogenetic invariants

Definition 2.1.1. An *algebraic variety* $\mathcal{V}$ in $\mathbb{C}^n$ is the set of solutions to a system of polynomial equations: $\mathcal{V} = \{p \in \mathbb{C}^n | f_1(p) = 0, \ldots, f_r(p) = 0\}$ for some polynomials $f_1, \ldots, f_r$ on n variables.
The set of algebraic varieties in $\mathbb{C}^n$ form the close sets of a topology, the *Zariski topology.*

Lemma 2.1.2. *Given any subset S of in $\mathbb{C}^n$ the set of polynomials vanishing on all the points in S forms an ideal $I(S)$ called ideal of S.*

Theorem 2.1.3. *Hilbert's Basis Theorem.*
Every ideal $I \subseteq \mathbb{C}[x_1, ..., x_n]$ can be generated by a finite set of polynomials $f_1, \ldots, f_m$.

Let $\mathcal{T}$ be a phylogenetic tree with n leaves and let T be its topology. Suppose that each node can take κ states from $\mathcal{K}$. Let $\mathcal{M}$ be an evolutionary model with d parameters on the tree topology, $p_{x_1,...,x_n}$ the joint distribution of nucleotides at the leaves defined as above and $\varphi_T^{\mathcal{M}} : \mathbb{C}^d \to \mathbb{C}^{4^n}$ the map that sends the set of d parameters to the set of the κ^n possible observations at the leaves.

The image $\mathrm{Im}_T^{\mathcal{M}}$ of the map $\varphi_T^{\mathcal{M}}$ contains the set of all the joint distributions of the states at the leaves for a tree topology T generated by some parameters in the evolutionary model $\mathcal{M}$. We denote by $\mathcal{V}_{\mathcal{M}}(T)$ the smallest algebraic variety containing $\mathrm{Im}_T^{\mathcal{M}}$ and we call it the *phylogenetic variety* associated to T and $\mathcal{M}$. The image set itself $\mathrm{Im}_T^{\mathcal{M}}$ is not in general an algebraic variety. But this set is a dense open subset in the smallest algebraic variety $\mathcal{V}_{\mathcal{M}}(T)$ containing it (in the Zariski topology).
We are going to study the ideal $I(\mathrm{Im}_T^{\mathcal{M}})$ (which coincides with the ideal of $\mathcal{V}_{\mathcal{M}}(T)$) and we will denote as $I_{\mathcal{M}}(T)$.

Definition 2.1.4. Given a tree topology T on n leaves and an evolutionary model $\mathcal{M}$, the polynomials in $I_{\mathcal{M}}(T)$ are called *invariants* of T. If f is a polynomial in $I_{\mathcal{M}}(T)$ which does not belong to $I_{\mathcal{M}(T')}$ for all the others tree topologies T' on n leaves, then f is called a phylogenetic invariant of T.

If a polynomial f lies on $I_{\mathcal{M}}(T)$, then we can observe that f defines a relationship among the probabilities $p_{x_1,...,x_n}$.

Example 2.1.5. This example will help us viewing a probabilistic model in phylogenetics algebraically. Consider a rally simple tree formed by a root r and two leaves to which we associate two random variables X_1 and X_2 (see Fig. 2.1).

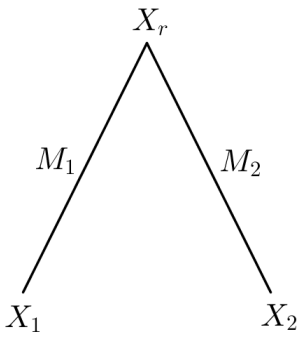

Figure 2.1: Statistical model on a rooted phylogenetic 2-leaf tree.

Suppose each the set of states of each random variable X_r, X_1, X_2 is $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. For the root r we specify the probabilities $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})$. For each edge of the tree we model the evolutionary process with two Markov matrices M_1 and M_2, where as we known M_i describes the mutation process on the edge e_i. Then from the model parameters we compute the probability of each possible observation at the leaves.

$$p_{x_1,x_2} = Prob(X_1 = x_1, X_2 = x_2) = \sum_{x_r \in \{A,C,G,T\}} \pi_{x_r} M_i(x_r, x_1) M_2(x_r, x_2). \tag{2.1}$$

The joint distribution $(p_{x_1,x_2})_{x_1,x_2}$ can be thought of as a 4×4 matrix each of whose entries is a polynomial of degree 3 and consisting of 4 terms in the parameters of the model. These 16 polynomials parameterize the model reflect all the modelling assumptions.

To produce a clear example and see some invariants we simplify the model by restricting it to an ancestral-A model. An ancestral-A model is such that the root sequence is composed of only the base $\mathtt{A}$ and then $\pi = (1, 0, 0, 0)$. Then the equation 2.1 is simpler and the joint distribution has a easier form

$$p_{x_1,x_2} = M_1(\mathtt{A}, x_1) M_2(\mathtt{A}, x_2). \tag{2.2}$$

In these conditions from equation 2.2 we observe

$$p_{x_1,x_2}p_{x_3,x_4} = M_1(\mathtt{A},x_1)M_2(\mathtt{A},x_2)M_1(\mathtt{A},x_3)M_2(\mathtt{A},x_4),$$

$$p_{x_1,x_4}p_{x_3,x_2} = M_1(\mathtt{A},x_1)M_2(\mathtt{A},x_4)M_1(\mathtt{A},x_3)M_2(\mathtt{A},x_2).$$

And therefore

$$p_{x_1,x_2}p_{x_3,x_4} - p_{x_1,x_4}p_{x_3,x_2} = 0.$$

Thus for every choice of x_1, x_2 and x_3, x_4 we have found a polynomial $f_{x_1x_2,x_3x_4}(P)$ that evaluate to 0 when $P = (p_{x_1,x_2})_{x_1,x_2}$ is any true distribution coming from the ancestral-A model.

$$f_{x_1x_2,x_3x_4}(P) = p_{x_1,x_2}p_{x_3,x_4} - p_{x_1,x_4}p_{x_3,x_2}.$$

Then we have found invariants for the ancestral-A model on a 2-leaves tree.

The problem that we have now is how can we found the generator basis of $I_{\mathcal{M}}(T)$? And if we know phylogenetic invariants of a tree, can we used to infer to the right topology?

For a concrete number of leaves, there exists some computational algebra programs that can calculate them using the kernel of $\varphi^T_{\mathcal{M}}$ (for example *singular* or *Macaulay2*). But in practice this is not even possible for 3-leaf trees, since this require so much memory capacity. Now we will mention some results that guarantee the possibility of calculate the invariants of n-leaf trees by the 3-leaf trees and the minors of some matrices.

Theorem 2.1.6. *([AR08], [Kut09]) Let $\mathcal{T}$ be a phylogenetic tree of n species $\mathcal{M}$ the model associated to this tree. There exists an algorithm to obtain a set of generators of $I_{\mathcal{M}}(\mathcal{T})$ from the invariants of a 3-leaf tree and the minors of certain matrices associated to the edges of T (we call them the edge invariants).*

This result it cannot be carried through, since if we pick GMM or the strand symmetric model, the invariants for 3-leaf trees are not known. The edge invariants are easy to compute and we will devote the next section to them. On the other hand it is not necessary to calculate a complete list of

invariants, for some applications we just need some of them. For example we just need to find the invariants that come from joint distributions and that define $V_{\mathcal{M}}(T)$. On the other hand if we are interested in just recovering the tree topology we just need the phylogenetic invariants. The following result shows us that the phylogenetic invariants are the edge invariants that we have mentioned above.

Theorem 2.1.7. *([CFS11]) Let $\mathcal{T}$ be a phylogenetic tree of n species evolving under an evolutionary $\mathcal{M}$. Then, for phylogenetic reconstruction purposes it is enough to consider only edge-invariants of T.*

2.2 Edge invariants for the general Markov model

To construct a first class of invariants we are going to consider the much simpler situation. Consider the next tree, two taxa a and b descendent from a common ancestor r, so the same tree as in the example 2.1.5. The Figure 2.1 shows the parameters associated to this tree. X_1, X_2, X_r are random variables associates to the nodes and M_1, M_2 are general Markov substitution matrices as above. Suppose $\kappa = 4$ and $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. We have calculated the joint distribution at the leaves at equation 2.1. Notice that there are 16 different probabilities since as x_1 as x_2 can have 4 values. Then we can represent these 16 expressions as a product of matrices:

$$P = (P_{x_1x_2})_{x_1x_2} = M_1^T diag(\pi) M_2,$$

where $diag(\pi)$ is a matrix that contains the vector π on the diagonal and 0 in all off-diagonal entries.

Now we will see some simple examples of how to find invariants based on the matrix P we have defined. We observe first the following easy lemma from linear algebra.

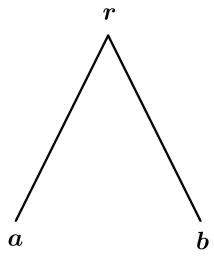

Figure 2.2: 2-leaved phylogenetic tree

Lemma 2.2.1. *Let $A_{m\times n}$ and $B_{n\times k}$ be $m \times n$ and $n \times k$ matrices. Then*

$$\mathrm{rk}(AB) \leq \min\{\mathrm{rk}\, A, \mathrm{rk}\, B\}.$$

Example 2.2.2. Ancestral-`A` model.
First of all consider the ancestral-`A` model that we have defined above, assuming the GM model with $\pi = (1, 0, 0, 0)$, so $diag(\pi)$ has only one non-zero entry, so has rank 1. This implies that the matrix P has rank 1 by Lemma 2.2.1. So in that case the 2×2 minors of P are all zero. These new equations are invariants for the ancestral-A model.

Example 2.2.3. Ancestral-`AC` model.
Now we define the ancestral-`AC` model on the 2 leaf tree as follows. Lets suppose $\pi = (\pi_A, 1 - \pi_A, 0, 0)$. Now diag($\pi$) has rank 2, and then the rank of P is at most 2. Thus all 3×3 minors are invariants.

Example 2.2.4. Ancestral-`ACG` model For the ancestral-`ACG` model, defined in the same way that the previous, i.e. $\pi = (\pi_A, \pi_C, 1 - \pi_A - \pi_C, 0)$ we can observe that P has at most rank 3, so $det(P) = 0$ is the only invariant that we can obtain with this method.

Example 2.2.5. For the General Markov model with no restrictions similar reasoning shows that P has at most rank 4, but P is already a 4×4 matrix,

so we do not obtain any invariant with this method.

To use this viewpoint to find invariants for the GMM we first need to generalize the definition of GMM to include the possibility of different number of states at the interior nodes and the leaves. In this case we allow the random variable X_i at node i to take values on a set of κ_i states and the transition matrix on edge e from node i to node j will be a $\kappa_i \times \kappa_j$ Markov matrix.

Lemma 2.2.6. *Let X_1, X_2, X_r be three discrete random variables associated to the nodes r, a and b of a 2-leaf tree in Figure 2.2. Suppose the variables may take κ, λ or μ states respectively. Let $P = (p_{x_1x_2})_{x_1x_2}$ where $p_{x_1x_2}$ is the probability of observing state x_1 at a and x_2 at b. Thus all $(\kappa+1)\times(\kappa+1)$ minors of P vanish. If $\kappa < \min\{\lambda, \mu\}$ these minors are invariants of the model.*

Proof. First of all, the vector π specifies the probabilities of the κ states at r while $M^1_{\kappa\times\lambda}$ and $M^2_{\kappa\times\mu}$ are the substitution matrices associates to the edges. We have already seen that $P = (M^1_{\kappa\times\lambda})^T diag(\pi) M^2_{\kappa\times\mu}$. We can observe that $diag(\pi)$ is a $\kappa\times\kappa$ matrix, and then its rank is at most κ, i.e. $\mathrm{rk}\,(diag(\pi)) \leq \kappa$. Thus, by Lemma 2.2.1 $\mathrm{rk}\,(P) \leq \kappa$ all $(\kappa+1)\times(\kappa+1)$ minors of P vanish. Obviusly, if $\kappa < \min\{\lambda,\mu\}$, then P is big enough to alllow room for such minors and therefore we have found some invariants for the model □

The following definition is crucial for finding phylogenetic invariants for the general Markov model.

Definition 2.2.7. Let $\mathcal{T}$ be a phylogenetic tree whith n leaves. Suppose that each random variable associated to a leaf can take κ states from $\mathcal{K} = \{x_1, \ldots, x_\kappa\}$. Let $A|B$ be a partition of the leaves (that is $L(\mathcal{T}) = A \cup B$ and $A \cap B = \emptyset$) and $\tilde{X}_A$ and $\tilde{X}_B$ the random variables associated to A and B. Then $\tilde{X}_A$ and $\tilde{X}_B$ can take $\kappa^{|A|}$ and $\kappa^{|B|}$ states respectively. We define the *flattening* $Flatt_{A|B}$ as a $\kappa^{|A|}\times\kappa^{|B|}$ matrix whose entries are the joint distribution at the leaves arranged according to the sets A and B:

$$Flatt_{A|B} = \begin{array}{c} \\ \textit{States at} \\ \textit{leaves} \\ \textit{in } A \end{array} \begin{array}{c} \textit{States at leaves in } B \\ \begin{pmatrix} p_{x_1x_1\ldots x_1x_1} & p_{x_1x_1\ldots x_1x_2} & p_{x_1x_1\ldots x_1x_3} & \cdots & p_{x_1x_1\ldots x_\kappa x_\kappa} \\ p_{x_1x_2\ldots x_1x_1} & p_{x_1x_2\ldots x_1x_2} & p_{x_1x_2\ldots x_1x_3} & \cdots & p_{x_1x_2\ldots x_\kappa x_\kappa} \\ p_{x_1x_3\ldots x_1x_1} & p_{x_1x_3\ldots x_1x_2} & p_{x_1x_3\ldots x_1x_3} & \cdots & p_{x_1x_3\ldots x_\kappa x_\kappa} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{x_\kappa x_\kappa\ldots x_1x_1} & p_{x_\kappa x_\kappa\ldots x_1x_2} & p_{x_\kappa x_\kappa\ldots x_1x_3} & \cdots & p_{x_\kappa x_\kappa\ldots x_\kappa x_\kappa} \end{pmatrix} \end{array}.$$

Example 2.2.8. Let $\mathcal{T}$ be a 4-leaf phylogenetic tree and $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Then $Flatt_{12|34}(P)$ is the 16×16 matrix:

$$Flatt_{A|B} = \begin{array}{c} \\ \textit{States at} \\ \textit{leaves} \\ 1 \textit{ and } 2 \end{array} \begin{array}{c} \textit{States at leaves } 3 \textit{ and } 4 \\ \begin{pmatrix} p_{\mathtt{AAAA}} & p_{\mathtt{AAAC}} & p_{\mathtt{AAAG}} & \cdots & p_{\mathtt{AATT}} \\ p_{\mathtt{ACAA}} & p_{\mathtt{ACAC}} & p_{\mathtt{ACAG}} & \cdots & p_{\mathtt{ACTT}} \\ p_{\mathtt{AGAA}} & p_{\mathtt{AGAC}} & p_{\mathtt{AGAG}} & \cdots & p_{\mathtt{AGTT}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{\mathtt{TTAA}} & p_{\mathtt{TTAC}} & p_{\mathtt{TTAG}} & \cdots & p_{\mathtt{TTTT}} \end{pmatrix} \end{array}.$$

As we mentioned in Section 1.4 we can view the vector of joint distribution p as a tensor in $\mathcal{W} \otimes \mathcal{W} \otimes \mathcal{W} \otimes \mathcal{W}$. Each component of this tensor product corresponds to one leaf, so in order to make leaves visible in this tensor product we denote it as $\mathcal{W}_1 \otimes \mathcal{W}_2 \otimes \mathcal{W}_3 \otimes \mathcal{W}_4$ ($\mathcal{W}_i = \mathcal{W}$). If we view the vector of joint distribution p as a tensor in $\mathcal{W}_1 \otimes \mathcal{W}_2 \otimes \mathcal{W}_3 \otimes \mathcal{W}_4$, then the flattening $Flatt_{12|34}(P)$ is the image of P via the isomorphism

$$\begin{array}{ccccc} \mathcal{W}_1 \otimes \mathcal{W}_2 \otimes \mathcal{W}_3 \otimes \mathcal{W}_4 & \cong & Hom(\mathcal{W}_1 \otimes \mathcal{W}_2, \mathcal{W}_3 \otimes \mathcal{W}_4) & \cong & M_{16\times 16} \\ p & & \mapsto & & Flatt_{12|34}(P) \end{array}$$

Here we prove the main theorem that gives invariants for the general Markov model.

Theorem 2.2.9. *Let $\mathcal{T}$ be the trivalent 4-leaf phylogenetic tree that has leaves 1,2 joined in a cherry. Suppose each random variable associated to a leaf can take states on a set $\mathcal{K}$ of cardinal κ. Then the $(\kappa+1)\times(\kappa+1)$ minors of $Flatt_{12|34}(P)$ vanish, equivalently $Flatt(P)$ has rank $\leq \kappa$.*

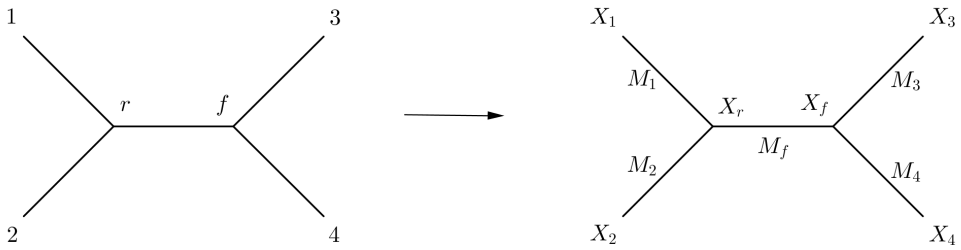

Figure 2.3: *Left:* 4-leaved tree, with taxa 1, 2, 3, 4, and rooted at r. *Right:* Statistical model on this 4-leaf tree.

Proof. Let $\mathcal{T}$ be a 4 leaves phylogenetic tree as the tree represented in Figure 2.3, with the root at the left of the internal edge. Every random variable X_i associated to the nodes can take κ states from a set $\mathcal{K}$. Then the GM model has as parameters a root distribution vector $\pi = (\pi_{x_1}, \ldots, \pi_{x_\kappa})$ and 5 $\kappa \times \kappa$ Markov matrices $M_1,\ M_2,\ M_3,\ M_4,\ M_f$. We can view the joint distribution P of this tree as a $\kappa \times \kappa \times \kappa \times \kappa$ array (X_1, X_2, X_3, X_4), where every entry is a random variable associated to a leaf.

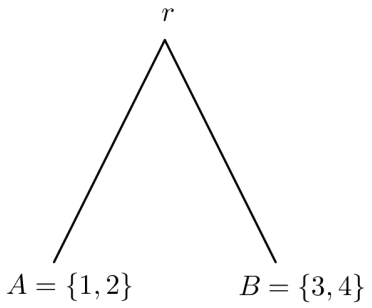

Figure 2.4: Split A={1,2}, B={3,4} in the 4-leaf tree in Figure 2.3

In order to use the previous lemma, we ignore some of the structure in the model by grouping the nodes as we can observe in Figure 2.4. Let $A = \{1,2\}, B = \{3,4\}$ be two pairs of nodes. The random variables associated to a and b, $\tilde{X}_1$ and $\tilde{X}_2$ respectively, have now κ^2 states.

For this model with the new structure we have the following Markov matrices $\tilde{M}_1$ and $\tilde{M}_2$, transition matrices from r to a and b.

$$\tilde{M}_1(x_i,(x_j,x_k)) = M_1(x_i,x_j)M_2(x_i,x_k),$$

$$\tilde{M}_2(x_i,(x_j,x_k)) = \sum_{l=1}^{\kappa} M_e(x_i,x_l)M_3(x_l,x_j)M_4(x_l,x_k).$$

The entries of $\tilde{M}_1$ are the probabilities that leaves 1 and 2 are in states x_j and x_k respectively if the root is in state x_i, and similary for $\tilde{M}_2$.

Use the GM model in this way corresponds tho changing the way we view the joint distribution array P. If we had a $\kappa \times \kappa \times \kappa \times \kappa$ array before, now we can consider one of the flattenings into a $\kappa^2 \times \kappa^2$ matrix. Notice that the entries of the array and the following matrix are unchanged, we just heve changed the way we view these entries.

$$Flatt_{A|B}(P)((i,j),(k,l)) = P_{i,j,k,l} = M_1^T diag(\pi) M_2,$$

$$Flatt(P)_{A|B} = \begin{pmatrix} p_{x_1x_1x_1x_1} & p_{x_1x_1x_1x_2} & p_{x_1x_1x_1x_3} & \cdots & p_{x_1x_1x_\kappa x_\kappa} \\ p_{x_1x_2x_1x_1} & p_{x_1x_2x_1x_2} & p_{x_1x_2x_1x_3} & \cdots & p_{x_1x_2x_\kappa x_\kappa} \\ p_{x_1x_3x_1x_1} & p_{x_1x_3x_1x_2} & p_{x_1x_3x_1x_3} & \cdots & p_{x_1x_3x_\kappa x_\kappa} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{x_\kappa x_\kappa x_1x_1} & p_{x_\kappa x_\kappa x_1x_2} & p_{x_\kappa x_\kappa x_1x_3} & \cdots & p_{x_\kappa x_\kappa \ldots x_\kappa x_\kappa} \end{pmatrix}$$

Thus this is a model for which we have already found invariants. We can therefore immediately see, using Lemma 2.2.6, that the $(\kappa+1)\times(\kappa+1)$ minors of $Flatt_{A|B}(P)$ are invariants of the of the GM model on this tree, since $Flatt_{A|B}(P)$ must have rank at most κ. □

It can be shown that for a dense subset of all parameters, the GM model with one specified root location on a tree T produces the same joint distribution as the GM model with a different root location on T, so these invariants do not depend on the location of r at one of the internal edge of the tree. Thus we can choose the root in the convenient location for our construction.

Remark 2.2.10. This construction easily generalizes to larger trees. We just have to pick an internal edge e joint two nodes r and f and we obtain a split $A|B$ by removing this edge of the tree. Then we have to construct the two matrices M_1 and M_2 in the same way that the latests ones. $\tilde{M}_1$ and $\tilde{M}_2$ will be $\kappa \times \kappa^{|A|}$ and $\kappa \times \kappa^{|B|}$ matrices, and their entries will be the conditional probabilities from r to the leaves of every subset of the partition. Then with the same argument we can see that the $(\kappa+1)\times(\kappa+1)$ minors of $Flatt_{A|B}(P)$ are invariants for this model.

Remark 2.2.11. Consequently, for any tree, considering the GMM with $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ the 5×5 minors of $Flatt_{A|B}(P)$ are invariants.

Notice that the entries in $Flatt_{A|B}(P)$, and thus the invariants we have found, depend only on the split of taxa $\{1,2\}$, $\{3,4\}$ induced by the internal edge of the tree. For larger trees we can pick any internal edge of T and we will obtain its particular $Flatt(P)$.

Example 2.2.12. Consider the 2-state GM model and a 5 leaves tree. Denoting states by 0 and 1. We have two possible splits, and we obtain two different flattenings (see Fig.2.5).

The $\{a_1, a_2, a_3\}, \{a_4, a_5\}$ split gives a 8×4 $Flatt_{a_1a_2a_3|a_4a_5}(P)$

$$Flatt_{a_1a_2a_3|a_4a_5}(P) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ \vdots & \vdots & \vdots & \vdots \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix},$$

and the $\{a_1, a_2\}, \{a_3, a_4, a_5\}$ split gives a 4×8 $Flatt_{a_1a_2|a_3a_4a_5}(P)$

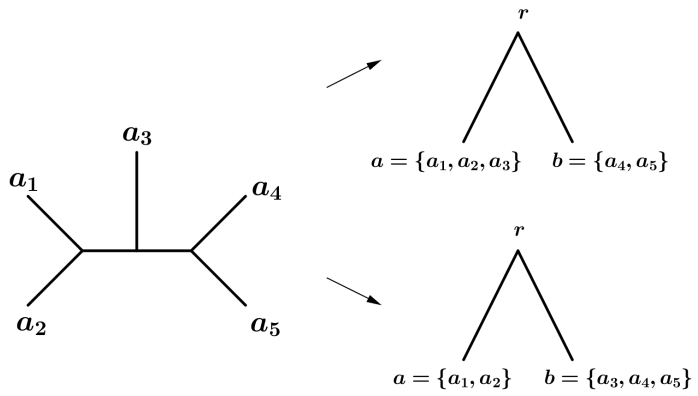

Figure 2.5: 5-leaf tree and two possible splits.

$$Flatt_{a_1a_2|a_3a_4a_5}(P) = \begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & \cdots & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & \cdots & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & \cdots & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & \cdots & p_{11111} \end{pmatrix}.$$

And therefore we will obtain different invariants from these two matrices.

We will prove now that the invariants we have already found are indeed *phylogenetic* invariants for 4-leaves tree. So the invariants we have found are not invariants in the other topologies. This means that the $(\kappa+1)\times(\kappa+1)$ minors of $Flatt_{12|34}(P)$ do not vanish if P is a distribution on the tree 13|24 or 14|23.

Theorem 2.2.13. *For general parameters on the tree* 12|34*,* $Flatt_{13|24}(P)$ *and* $Flatt_{14|23}(P)$ *have rank* κ^2*.*

Proof. Consider the four leaves tree as in Figure 2.3, a κ-state GM model and $\mathcal{K}$ the states of each random variable X_i associated to the node i. Consider now the partition $A = \{1,3\}$ and $B = \{2,4\}$ (see Figure 2.6). The

joint distrubutions have not changed, but now M_A (transition matrix from r to A), M_B (transition matrix from r to B) and Flatt(P) have new structures. We have seen that $rank(Flatt_{12|34}(P)) \leq 4$ we will prove now that $rank(Flatt_{13|24}(P)) = 16$ if P is general enough.

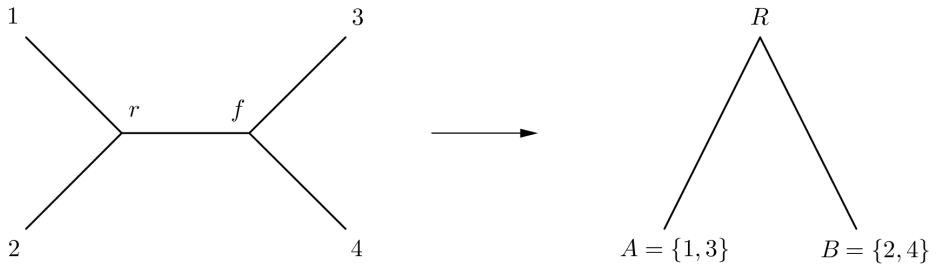

Figure 2.6: *Left:* 4-leaved tree *Right:* Split A={1,3}, B={2,4}

Let R be the set of the nodes that are shared by the induced subtrees for a and b. In this case $R = \{r, f\}$. Then we can write

$$Flatt_{A|B}(P) = M_A^T diag(\pi(R)) M_B$$

where $\pi(R)$ is the distribution of R and M_A and M_B are the $\kappa^{|R|} \times \kappa^{|A|}$ and $\kappa^{|R|} \times \kappa^{|B|}$ transition matrices. In our case, $|R| = |A| = |B| = 2$. So $\pi(R)$, M_A and M_B are $\kappa^2 \times \kappa^2$ matrices.

$\pi(R)$ is a diagonal matrix, and the entries are the probabilities of the κ^2 possible states for $\{r, f\}$. Furthermore $rank(\pi(R)) = \kappa^2$ if $\pi(R)$ has no zero entries (in this step we make use of the fact that parameters are general).

Let us see how is M_A structured.

$$M_A = (P(A = x_i x_j | R = x_u x_v))_{\substack{x_i x_j \in \mathcal{K}^2 \\ x_u x_v \in \mathcal{K}^2}} = (P(1 = x_i, 3 = x_j | r = x_u, f = x_v))_{x_i, x_j, x_u, x_v \in \mathcal{K}}$$

Therefore

$$P(a = x_i x_j | R = x_u x_v) = P(1 = x_i | r = x_u) P(3 = x_j | f = x_v) = M_1(x_u, x_i) M_3(x_v, x_j)$$

So we can deduce that

$$M_A = (M_1 \otimes M_3)^T$$

Thus we have proved that M_A is a Kronecker product[1], and

$$\mathrm{rk}(M_1 \otimes M_3) = \mathrm{rk}(M_1)\,\mathrm{rk}(M_3).$$

The matrix M_B is constructed similarly. If M_1 and M_3 are general stocasthic matrices, $\mathrm{rk}(M_i) = \kappa$ so $\mathrm{rk}(M_1 \otimes M_3) = \kappa^2$. Consequently

$$Flatt_{A|B}(P) = M_A^T diag(\pi(R)) M_B = (M_1 \otimes M_3)^T diag(\pi(R))(M_2 \otimes M_4).$$

And $rank(Flatt_{A|B}(P))$ is at most κ^2 since $rank(M_A) = rank(diag(\pi(R))) = rank(M_B) = \kappa^2$.

□

Example 2.2.14. Suppose that now $\kappa = 4$ and $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ Then we will see the structure of M_a in this case.

$$M_a = \begin{pmatrix} p_{\mathtt{AAAA}} & p_{\mathtt{ACAA}} & \cdots & p_{\mathtt{TTAA}} \\ p_{\mathtt{AAAC}} & p_{\mathtt{ACAC}} & \cdots & p_{\mathtt{TTAC}} \\ \vdots & \vdots & \ddots & \vdots \\ p_{\mathtt{AATT}} & p_{\mathtt{ACTT}} & \cdots & p_{\mathtt{TTTT}} \end{pmatrix}$$

And we observe that these probabilities can be written as $p_{\mathtt{ACGT}} = Prob(a_1 = \mathtt{A}, a_3 = \mathtt{C} | r = \mathtt{G}, f = \mathtt{T}) = Prob(a_1 = \mathtt{A} | r = \mathtt{G}) Prob(A_3 = \mathtt{C} | f = \mathtt{T}) = M_1(\mathtt{G}, \mathtt{A}) M_3(\mathtt{T}, \mathtt{C})$. Therefore

[1] If A is an $m \times n$ matrix and B is a $p \times q$ matrix, then the *Kronecker product* $A \otimes B$ is the $mp \times nq$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \dots & a_{nn}B \end{pmatrix}.$$

$$M_A = \begin{pmatrix} M_1(\mathtt{A},\mathtt{A})M_3(\mathtt{A},\mathtt{A}) & M_1(\mathtt{A},\mathtt{A})M_3(\mathtt{A},\mathtt{C}) & \dots & M_1(\mathtt{A},\mathtt{T})M_3(\mathtt{A},\mathtt{T}) \\ M_1(\mathtt{A},\mathtt{A})M_3(\mathtt{C},\mathtt{A}) & M_1(\mathtt{A},\mathtt{A})M_3(\mathtt{C},\mathtt{C}) & \dots & M_1(\mathtt{A},\mathtt{T})M_3(\mathtt{C},\mathtt{T}) \\ \vdots & \vdots & \ddots & \vdots \\ M_1(\mathtt{T},\mathtt{A})M_3(\mathtt{T},\mathtt{A}) & M_1(\mathtt{T},\mathtt{A})M_3(\mathtt{T},\mathtt{C}) & \dots & M_1(\mathtt{T},\mathtt{T})M_3(\mathtt{T},\mathtt{T}) \end{pmatrix} =$$

$$= \begin{pmatrix} M_1(\mathtt{A},\mathtt{A})M_3 & M_1(\mathtt{A},\mathtt{C})M_3 & M_1(\mathtt{A},\mathtt{G})M_3 & M_1(\mathtt{A},\mathtt{T})M_3 \\ M_1(\mathtt{C},\mathtt{A})M_3 & M_1(\mathtt{C},\mathtt{C})M_3 & M_1(\mathtt{C},\mathtt{G})M_3 & M_1(\mathtt{C},\mathtt{T})M_3 \\ M_1(\mathtt{G},\mathtt{A})M_3 & M_1(\mathtt{G},\mathtt{C})M_3 & M_1(\mathtt{G},\mathtt{G})M_3 & M_1(\mathtt{G},\mathtt{T})M_3 \\ M_1(\mathtt{T},\mathtt{A})M_3 & M_1(\mathtt{T},\mathtt{C})M_3 & M_1(\mathtt{T},\mathtt{G})M_3 & M_1(\mathtt{T},\mathtt{T})M_3 \end{pmatrix} =$$

$$= M_1 \otimes M_3.$$

Example 2.2.15. We have tried to verify Theorem 2.2.13 with a concrete example. We have picked 5 random matrices $M_1,\ M_2,\ M_3,\ M_4,\ M_f$ (see Fig. 2.3) and a random vector π.

$$\pi = (0.22, 0.26, 0.24, 0.28)$$

$$M_1 = \begin{pmatrix} 0.7 & 0.15 & 0.10 & 0.05 \\ 0.07 & 0.75 & 0.16 & 0.02 \\ 0.12 & 0.08 & 0.68 & 0.12 \\ 0.05 & 0.08 & 0.07 & 0.8 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 0.82 & 0.05 & 0.12 & 0.01 \\ 0.11 & 0.6 & 0.07 & 0.22 \\ 0.07 & 0.14 & 0.75 & 0.04 \\ 0.12 & 0.14 & 0.10 & 0.64 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 0.67 & 0.11 & 0.09 & 0.13 \\ 0.06 & 0.75 & 0.14 & 0.05 \\ 0.07 & 0.15 & 0.63 & 0.15 \\ 0.04 & 0.08 & 0.16 & 0.72 \end{pmatrix}$$

$$M_4 = \begin{pmatrix} 0.71 & 0.13 & 0.1 & 0.06 \\ 0.13 & 0.63 & 0.14 & 0.10 \\ 0.12 & 0.06 & 0.80 & 0.02 \\ 0.09 & 0.09 & 0.11 & 0.77 \end{pmatrix}$$

$$M_f = \begin{pmatrix} 0.59 & 0.16 & 0.12 & 0.13 \\ 0.12 & 0.66 & 0.08 & 0.14 \\ 0.07 & 0.16 & 0.73 & 0.04 \\ 0.18 & 0.10 & 0.08 & 0.64 \end{pmatrix}$$

We have calculated the joint distributions $P = (p_{\texttt{AAAA}}, p_{\texttt{AAAC}}, \ldots, p_{\texttt{TTTT}})$ for 4-leaves tree 12|34 represented at Fig. 2.3. Then we have evaluated this probabilities in the matrices $Flatt_{12|34}(P), Flatt_{13|24}(P), Flatt_{14|32}(P)$ and we have calculatet their ranks. The results have been

$$\operatorname{rk} Flatt_{12|34}(P) = 4$$

$$\operatorname{rk} Flatt_{13|24}(P) = 16$$

$$\operatorname{rk} Flatt_{14|32}(P) = 16$$

See Annex A to inspect the c++ code that calculates the joint distributions, flattening matrices and their ranks.

Appendix A

Computation of rank of flattening matrices

```
#include <iostream>
#include <fstream>
#include <vector>
#include <algorithm>
#include <cmath>

using namespace std;

typedef vector <double> VE;
typedef vector < VE > ME;
const double eps = 1e-6;
VE pi;
ME M1;
ME M2;
ME M3;
ME M4;
ME Mf;

void read_matrix(M &M){
```

```
   for (int i = 0; i < 4; ++i){
      for (int j = 0; j < 4; ++j) {
         cin >> M[i][j];
      }
   }
}

void write_matrix(ME &M){
   for (int i = 0; i < (int)M.size(); ++i){
      for (int j = 0; j < (int) M[i].size(); ++j) cout << M[i][j] << " ";
         cout << endl;
      }
   cout << endl;
}

double joint_distribution(int a, int b, int c, int d){
   double Paaaa = 0;

   for(int i = 0; i < 4; ++i){
      double sum = 0;
      for (int j = 0; j < 4; ++j)
         sum = sum + Mf[i][j]*M3[j][c]*M4[j][d];
      Paaaa = Paaaa + pi[0]*M1[i][a]*M2[i][b]*sum;
   }
   return Paaaa;
}

void clean (ME &M){
   for (int i = 0; i < (int) M.size(); ++i){
      for (int j = 0; j < (int) M[i].size(); ++j) {
         if (abs(M[i][j]) < eps) M[i][j] = 0;
      }
```

```
   }
}

void Gauss(ME &M){
   int files = 0;
   int i = 0;
   while (files < (int)M.size() and i < (int)M[0].size()){
      if (abs(M[files][i]) < eps){
         bool trobat = false;
         for (int j = files+1; j < (int)M.size() and not trobat; ++j){
            if (abs(M[j][i]) > eps){
               trobat = true;
               swap(M[j], M[files]);
            }
         }
      }
      if (abs(M[files][i]) > eps){
         for (int j = i+1; j < (int)M.size(); ++j){
            if (abs(M[j][i]) > eps){
               double pivot = M[j][i];
               for (int k = i; k < (int)M[0].size(); ++k)
                  M[j][k] -= (M[files][k]*pivot)/M[files][i];
               }
            }
         ++files;
      }
      ++i;
   }
   clean(M);
}

int rank (ME &M){
```

```
   int cont = 0;
   for (int i = 0; i < (int) M.size(); ++i)
      if (abs(M[i][i]) > eps) ++cont;
   return cont;
}

int main(){

   pi = VE(4);   M1 = ME(4, VE(4));   M2 = ME(4, VE(4));
   M3 = ME(4, VE(4));
   M4 = ME(4, VE(4));
   Mf = ME(4, VE(4));

   for (int i = 0; i < 4; ++i) cin >> pi[i];
   read_matrix(M1);
   read_matrix(M2);
   read_matrix(M3);
   read_matrix(M4);
   read_matrix(Mf);

   ME Flatt12_34 (16, VE(16));
   ME Flatt13_24 (16, VE(16));
      ME Flatt14_23 (16, VE(16));

   int a; int b; int c; int d;

   for (int i = 0; i < 16; ++i){
      for (int j = 0; j < 16; ++j){
         a = i/4;
         b = i%4;
         c = j/4;
         d = j%4;
```

```
            Flatt12_34 [i][j] = joint_distribution (a, b, c, d);
        }
    }

    for (int i = 0; i < 16; ++i){
        for (int j = 0; j < 16; ++j){
            a = i/4;
            b = j/4;
            c = i%4;
            d = j%4;
            Flatt13_24[i][j] = joint_distribution (a, b, c, d);
        }
    }
    for (int i = 0; i < 16; ++i){
        for (int j = 0; j < 16; ++j){
            a = i/4;
            b = j/4;
            c = j%4;
            d = i%4;
            Flatt14\_23[i][j] = joint_distribution (a, b, c, d);
        }
    }

    Gauss(Flatt12_34); write\_matrix(Flatt12_34);
    Gauss(Flatt13_24); write\_matrix(Flatt13_24);
    Gauss(Flatt14_23); write\_matrix(Flatt14_23);

    cout << "The matrix Flatt(12|34) has rank " << rank(Flatt12_34) << endl;
    cout << "The matrix Flatt(13|24) has rank " << rank(Flatt13_24) << endl;
    cout << "The matrix Flatt(14|23) has rank " << rank(Flatt14_23) << endl;
}
```

References

[AR04] ES Allman and JA Rhodes. *Mathematical models in biology, an introduction*. Cambridge University Press, January 2004. ISBN 0-521-52586-1).

[AR05] ES Allman and JA Rhodes. The mathematics of phylogenetics. University of Alaska Fairbanks, 2005.

[AR07] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants. In *Reconstructing evolution*, pages 108–146. Oxford Univ. Press, Oxford, 2007.

[AR08] ES Allman and JA Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Advances in Applied Mathematics*, 40:127–148, 2008.

[Cas12] M Casanellas. Algebraic tools for evolutionary biology. *La Gaceta de la RSME*, 15:521–536, 20012.

[CFS10] M. Casanellas and J. Fernandez-Sanchez. Reconstrucción filogenética usando geometría algebraica. *Arbor. Ciencia, pensamiento, cultura*, 96:207–229, 2010.

[CFS11] M Casanellas and J Fernandez-Sanchez. Relevant phylogenetic invariants of evolutionary models. *Journal de Mathématiques Pures et Appliquées*, 96:207–229, 2011.

[Eri05] N Eriksson. Tree construction using singular value decomposition. In L Pachter and B Sturmfels, editors, *Algebraic Statistics for computational biology*, chapter 19, pages 347–358. Cambridge University Press, 2005.

[Kut09] J. Draisma J. Kuttler. *On the ideals of equivariant tree models*, volume 344 of *Mathematische Annalen*. Springer, 2009.