# Phylogenetic Reconstruction Based on Algebra

Marta Casanellas(✉)

Universitat Politècnica de Catalunya, Diagonal 628, Barcelona, Spain
marta.casanellas@upc.edu

**Abstract.** Phylogenetics is the discipline that studies the evolutionary history of species. During the last years there has been an approach to phylogenetics from the point of view of algebraic geometry. This perspective has been used to study the evolutionary models most used in phylogenetics but also to develop new phylogenetic reconstruction tools. Here we review the interplay between algebra and phylogenetics and we explain the most recent results that use these methods for the purpose of phylogenetic reconstruction.

## 1 Introduction

According to Darwin's theory of natural selection, the evolution of species is usually represented on a *phylogenetic tree*: its leaves represent living species, the interior nodes represent their common ancestors, and edges represent evolutionary processes (see Fig. 1). Nowadays, the study of the evolutionary history of a group of species is carried out from deoxyribonucleic acid (DNA) molecules associated to the living species in the study. These DNA molecules may correspond to certain genes or to other molecular entities and, due to the double helix structure of DNA, they can be seen as words on the alphabet $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ with $\mathtt{A}$ denoting adenine, $\mathtt{C}$ cytosine, $\mathtt{G}$ guanine, and $\mathtt{T}$ thymine. In this sense they are called *DNA sequences*. The aim of phylogenetics is to reconstruct the ancestral relationships among species (i.e. the phylogenetic tree) from a given set of DNA sequences.

Phylogenetics is not only aimed at the knowledge of evolutionary history per se: it has applications on many different topics nowadays. For example, as pointed out in [3], phylogenetics is used in the design of biodiversity preservation policies, in the prediction of molecular evolution of viruses, or in the detection of tumor origins. It has recently been used in determining the origin of SARS-CoV-2 and the phylogenetic tree of coronaviruses has been crucial in detecting the relationship between this coronavirus in bats and humans, see [4]. Phylogenetics has also an impact beyond biology: it is a crucial tool in the study of origin and evolution of languages and written texts, for instance.


H. Herrero et al. (Eds.): RSME 2022, RSME 13, pp. 1–19, 2024.
https://doi.org/10.1007/978-3-031-62025-6_2

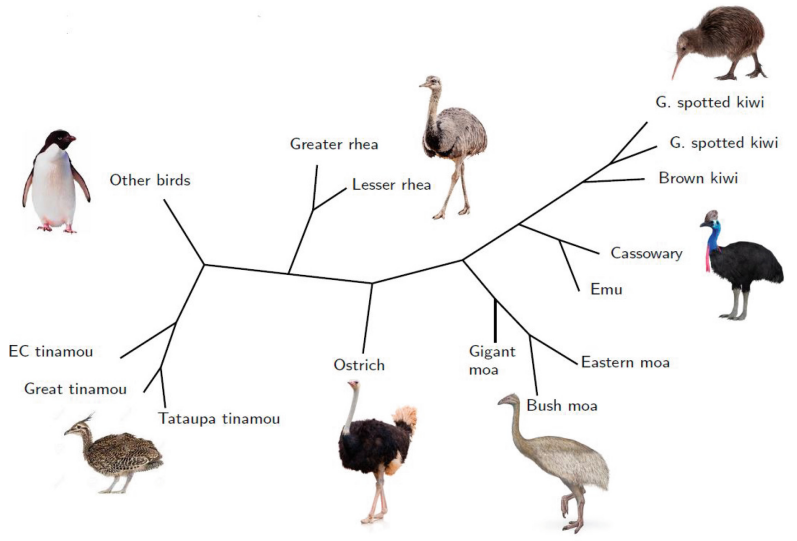


**Fig. 1.** A phylogenetic tree on several species of birds. Rheas, Kiwi, Ostrich, Cassowary and Emus, (extinct) Moas and Tinamous are paleognathus birds; they are known as ratites and among them only Tinamous can fly. It is still controversial whether this would be the correct phylogeney of these species, or alternatively, Tinamous should be placed within the clade of other Ratites, see [1,2]. Figure courtesy of Marina Garrote-López.

The main goal in phlyogenetic reconstruction is the following:

**? Problem: phylogenetic reconstruction**

Provide consistent methods that, given a collection of DNA sequences, produce the most plausible phylogenetic tree representing the evolution of the sequences.

Nowadays phylogenetics faces many different challenges. As there are more and more data available, there is a need of using more complex models fitting these data. On the contrary, phylogenetic reconstruction tools usually assume oversimplified models to make computations feasible. Moreover, the number of phylogenetic trees grows more than exponentially in the number of leaves (or species under study), so it is not possible to explore exhaustively the space of all phylogenetic trees. Thus, there is a need for new methods working for more complex models.

Here, the word "plausible" has a vague meaning on purpose, as it depends on the measure one wants to use. In order to reconstruct the phylogenetic tree, one usually models the evolution via a Markov process on the tree. This Markov process governs the substitution of nucleotides along the tree and is the basis of widely used methods such as maximum likelihood or Bayesian tools. At the beginning of this century, the emergence of the new discipline *algebraic statistics* (coined as in the book [5]) made possible the use of algebraic tools in statistical inference. These tools have been used in computational biology since the papers of L. Pachter and B. Sturmfels at Proceedings of the National Academy of Sciences, [6] and [7], and appear in many different areas nowadays.

In phylogenetics, the use of algebraic tools was initiated by E. Allman and J. Rhodes, and at present there is a solid community working in *algebraic phylogenetics*. The key observation is that Markov processes on trees are algebraic statistical models (i.e. parametric models for which the distribution is expressed as a polynomial in the parameters). From this, the study of algebraic varieties they define appears naturally. This has lead to relevant consequences from algebraic geometry in phylogenetics (see [8,9]) but also problems from phylogenetics have lead to important results in mathematics (see [10–13]) and other areas, see [14]. Recently, a refinement has been introduced in order to use semi-algebraic varieties instead of algebraic varieties, [15–17].

In this article we review the most important algebraic tools that are used in phylogenetic reconstruction. In Sect. 2 we give the main definitions, describe Markov processes on phylogenetic trees and how these lead to the natural study of the corresponding algebraic varieties. In Sect. 3 we provide the main equations that define these algebraic varieties. In Sect. 4 we state the semi-algebraic conditions that must be taken into account in phylogenetics. Finally, in Sect. 5 we explain how to use these tools in phylogenetic reconstruction. This is based on previous joint work with J. Fernández-Sánchez and M. Garrote-López in [17] and [16].

## 2 Markov Processes of Nucleotide Substitution

### 2.1 Phylogenetic Trees

We start with the basic definitions of phylogenetic trees, which can be found in the book by Mike Steel [18, chapter 1] for example.

A *tree* $T$ is a connected acyclic graph with a collection of vertices $V(T)$ and edges $E(T)$. The *degree* of a vertex $v \in V(T)$ is the number of edges that are incident with $v$. The set of vertices of degree 1 are called *leaves* and those non-leaf vertices are called *interior* vertices.

**Definition 1.** *Let $S$ be a finite set (in our setting $S$ is usually a set of biological entities such as living species). An* (unrooted) phylogenetic tree on $S$ *is a tree $T$ with leaf set $S$ whose interior vertices have degree at least three.*

In other words, a phylogenetic tree on $S$ is a tree with no nodes of degree two together with a bijection between the set of leaves of $T$ and the set $S$. The set $S$ represents current species while the interior nodes represent their common ancestors, see Fig. 1 for an example. The set $S$ will be usually understood from the context, so it will be often omitted.

Two phylogenetic trees $T$ and $T'$ on $S$ are *isomorphic* if there is a graph isomorphism from $T$ to $T'$ that is the identity on the leaf set $S$. The topology of a phylogenetic tree refers to the isomorphism class of $T$ as phylogenetic tree. See Fig. 2 for three classes of isomorphism of phylogenetic trees on $S = \{1, 2, 3, 4\}$, that will be represented as $T_{12|34}$, $T_{13|24}$, $T_{14|23}$.

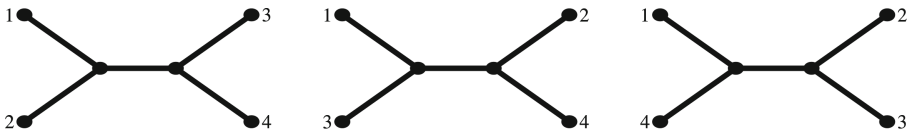


**Fig. 2.** Three non-isomorphic trees on the set of leaves $1, 2, 3, 4$. These trees are denoted as $T_{12|34}$, $T_{13|24}$, and $T_{14|23}$ (from left to right) and, together with the star tree on four leaves, they represent all possible tree topologies on $1, 2, 3, 4$.

We can *root* a phylogenetic tree by specifying a vertex $r$ in $V(T)$ and directing all edges out of it. For reasons related to identifiability that will be clarified in the next subsection, in our setting we do not allow trees with degree two nodes.

There is another type of information that can be added on a phylogenetic tree: if there are weights assigned to the edges (or *branch lengths*), these usually represent an evolutionary distance between both ends of the edge. In this paper we do not take into account this kind of information.

### 2.2 Models of Nucleotide Substitution

In order to study the evolutionary relationships among DNA sequences, one specifies a mathematical model of substitution of nucleotides along a phylogenetic tree. Let $T$ be an unrooted phylogenetic tree with leaves $S = \{1, \ldots, n\}$ and set an internal vertex $r$ to play the role of the root so that we can direct the edges out of it (see for example, Fig. 3). The vertex $r$ had a DNA sequence associated to it (a word on the alphabet `A`, `C`, `G`, `T`) and this sequence has randomly mutated to the DNA sequences that we observe nowadays at the leaves of the tree. To simplify things, it is common to assume that the nucleotides at different positions evolve independently of each other and following the same process (that is, sites at the DNA sequence are independent and identically distributed). Thus one only needs to model the evolution of a single nucleotide.

Assign a random variable $X_i$ at each node $i \in V(T)$ and assume that the substitution of nucleotides along the tree follows a Markov process: the random variable at each node is conditionally independent of its non-descendant random

variables given the random variable at its immediate parent node. Another way to describe this Markov process is via a parametric statistical model: let $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})$ be the distribution of nucleotides at the root of a tree ($\pi_{\mathtt{X}}$ is the probability of $\mathtt{X}$ at the DNA sequence at $r$) and let $M^e$ be the matrix of conditional probabilities of substitution along edge $e : u \longrightarrow v$, that is,

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

Here $P(\mathtt{X}|\mathtt{Y}, e)$ denotes the conditional probability that nucleotide $\mathtt{Y}$ at the parent node $u$ of $e$ is substituted by nucleotide $\mathtt{X}$ at the child node $v$. Note that this matrix has non-negative entries and sum of rows equal to one; this is called a Markov matrix, transition matrix, or row stochastic matrix. With these parameters, the Markov process on the tree is specified by

$$Prob\left(\{\mathcal{X}_w = \mathtt{X}_w\}_{w \in V(T)}\right) = \pi_{\mathtt{X}_r} \prod_{e:u \to v} M^e_{\mathtt{X}_u, \mathtt{X}_v}.$$

On a phylogenetic tree we do not have observations of the ancestral DNA sequences, so the random variables at the interior nodes of the tree are hidden. Thus, we obtain the probability $p_{\mathtt{X}_1 \ldots \mathtt{X}_n}$ of observing nucleotides $\mathtt{X}_i$ at leaf $i$ by marginalizing the previous expression over the interior nodes:

$$p_{\mathtt{X}_1 \ldots \mathtt{X}_n} = \sum_{\substack{\mathtt{X}_u \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\} \\ u \in Int(T)}} \pi_{\mathtt{X}_r} \prod_{e:u \to v} M^e_{\mathtt{X}_u, \mathtt{X}_v}, \tag{1}$$

where $Int(T)$ denotes the set of interior nodes of $T$.

For example, for the phylogenetic tree of Fig. 3, we have

$$p_{\mathtt{ACCG}} = \sum_{\mathtt{X}_r} \sum_{\mathtt{X}_u} \pi_{\mathtt{X}_r} M^1_{\mathtt{X}_r, \mathtt{A}} M^2_{\mathtt{X}_r, \mathtt{C}} M^5_{\mathtt{X}_r, \mathtt{X}_u} M^3_{\mathtt{X}_u, \mathtt{C}} M^4_{\mathtt{X}_u, \mathtt{G}}.$$

The entries of $\pi$ and $M^e$ are parameters of this statistical model and, as we have just seen, the joint distribution of nucleotides at the leaves of the tree can be expressed as a polynomial function of these parameters. Therefore we have an algebraic statistical model and the following polynomial map sends each set of free parameters to the joint distribution $p^T = (p_{\mathtt{X}_1 \ldots \mathtt{X}_n})_{\mathtt{X}_1, \ldots, \mathtt{X}_n}$ of nucleotides at the leaves:

$$\begin{array}{c} \varphi_T : \text{Free Parameters} \longrightarrow \mathbb{R}^{4^n} \\ \left(\text{entries of } \pi, (M^e)_{e \in E(T)}\right) \mapsto p^T = (p_{\mathtt{AA}\ldots\mathtt{A}}, p_{\mathtt{AA}\ldots\mathtt{C}}, p_{\mathtt{AA}\ldots\mathtt{G}}, \ldots, p_{\mathtt{TT}\ldots\mathtt{T}}). \end{array} \tag{2}$$

Any distribution $p$ arising from a Markov process on the tree $T$ is in the image of this map. Characterizing which distributions are in the image of this

map is crucial for deciding whether a given distribution from real data is likely to have arisen as a Markov process on $T$ or not. We come back to this problem in the next subsection, but for the moment we want to mention other evolutionary models.

The model presented above is commonly known as the *general Markov* (GM briefly) model, or the Barry-Hartigan model [19]. Depending on the biological data we are dealing with, simpler models can be considered. In what follows, we describe these simpler models as instances of *equivariant* models (see [20]).

Let $G$ be a subgroup of the symmetric group $\mathfrak{S}_4$ on the set $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$, i.e. $G$ is a group of permutations of these four elements. If in the parametric statistical model presented above we impose

(1) the distribution $\pi$ is invariant by the action of $G$, that is, $(\pi_{\mathtt{A}},\pi_{\mathtt{C}},\pi_{\mathtt{G}},\pi_{\mathtt{T}}) = (\pi_{g\mathtt{A}},\pi_{g\mathtt{C}},\pi_{g\mathtt{G}},\pi_{g\mathtt{T}})$ for any $g \in G$, and
(2) each transition matrix $M^e$ is a $G$-equivariant map: $M^e_{\mathtt{X},\mathtt{Y}} = M^e_{g\mathtt{X},g\mathtt{Y}}$ for any $g \in G$,

then we have a $G$-*equivariant* model of nucleotide substitution. Well known examples of $G$-equivariant models are (listed in decreasing order of complexity):

- The GM model above: it is a $G$-equivariant model if we take $G = \{id\}$. In this model we have 12 free parameters per edge (for each transition matrix) plus three free parameters for the distribution $\pi$ at the root. The GM model evolving on the tree of Fig. 3 has 63 free parameters.
- Strand-symmetric model, see [21]: if $G = \langle(\mathtt{AT})(\mathtt{CG})\rangle$, the corresponding equivariant model preserves the symmetry between both strands of DNA molecules. In this case, as the distribution $\pi$ at the root must be $G$-invariant, it satisfies $\pi_{\mathtt{A}} = \pi_{\mathtt{T}}$ and $\pi_{\mathtt{C}} = \pi_{\mathtt{G}}$.
- Kimura 3-parameter model, see [22]: when $G = \langle(\mathtt{AC})(\mathtt{GT}),(\mathtt{AG})(\mathtt{CT})\rangle$ the distribution at the root must be uniform $\pi = (\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$ and the transition matrices satisfy symmetries that reflect the chemical properties of both groups of nucleotides, purines (adenine $\mathtt{A}$ and guanine $\mathtt{G}$) and pyrimidines (cytosine $\mathtt{C}$ and thymine $\mathtt{T}$). For this model and its submodels below, the uniform distribution $\pi$ is the stationary distribution of all transition matrices of the model.
- Kimura 2-parameters, abbreviated as K80 (see [23]): $G = \langle(\mathtt{ACGT}),(\mathtt{AG})\rangle$; this is a submodel of K81 that considers all substitutions between purines and pyrimidines to be equally probably at each edge.

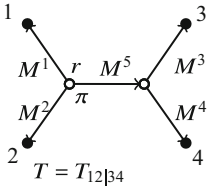


**Fig. 3.** A Markov process on the tree $T_{12|34}$.

- Jukes-Cantor model, abbreviated as JC69 (see [24]): this is the most simple model and corresponds to $G = \mathfrak{S}_4$. In this case the distribution at the root is uniform and the transition matrices at the edges are of the following type:

$$M^e = \begin{pmatrix} a_e & b_e & b_e & b_e \\ b_e & a_e & b_e & b_e \\ b_e & b_e & a_e & b_e \\ b_e & b_e & b_e & a_e \end{pmatrix},$$

  where $a_e + 3b_e = 1$. For this model there is only one free parameter per edge (and there are no free parameters for the distribution at the root), thus this model on the tree of Fig. 3 has only five free parameters.

One of the first issues that one has to take into account when proposing a parametric model is *identifiability*, that is, whether all parameters of the model can be identified from data that perfectly fits the model. In our setting this translates to the following questions:

**? Questions on identifiability**

Let $p$ be a distribution that has arisen as a Markov process on a phylogenetic tree $T$ with parameters $\pi$ and $M^e$.

- Can the tree $T$ be identified solely from $p$?
- If the answer to the previous question is yes, can the parameters $\pi$ and $M^e$ be uniquely identified from $p$?

---

It is well known (see [18, §7.2.1]) that the answer to the first question is affirmative if $p$ has arisen from *non-singular* parameters (that is, $\pi$ has no zero entries and all matrices $M^e$ are invertible and different from permutation matrices), which is a generic condition. Obviously, $T$ can be recovered only up to isomorphism (so it is actually the tree topology that can be recovered from $p$); however, the vertex chosen to root the tree is not identifiable.

Furthermore, under the same conditions that guarantee an affirmative answer to the first question, there is an affirmative answer for the second: if we fix an interior vertex $r$ to direct the tree, $\pi$ and all transition matrices $M^e$ can be recovered from $p$ up to label swapping of states $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ at the interior nodes (see [25] for the GM model and [11] for the other $G$-equivariant models).

Thus the Markov models presented above are well-posed, in the sense that there is no overparametrization. Of course there are some sets of parameters that induce distributions at the leaves that can arise on differen trees, but this only occurs for particular parameters as we have seen (for example, in the tree of Fig. 3 we can set $M^5 = id$ and then the corresponding joint distribution at the leaves can also be obtained as a Markov process on any of the other two trees of Fig. 2).

If we had allowed nodes of degree two, then the answer to the identifiability questions would have been negative: if there is a node $v$ of degree two (as the root in a binary rooted tree), one can only recover the product of the transition matrices of both edges incident with $v$ (not the matrices separately).

There are more complex models that could be considered (for example allowing different sites to evolve according to different processes or allowing insertions and deletions of nucleotides and not only substitutions), but this is out of the scope of this survey.

### 2.3 Algebraic Geometry in Phylogenetics

If $\Delta$ is the standard simplex in $\mathbb{R}^{4^n}$, characterizing which distributions $p \in \Delta$ arise as a Markov process on a phylogenetic tree $T$ can be useful in recovering the tree $T$ from a given distribution. This is the main idea that leads to the use of algebraic geometry for phylogenetic reconstruction. The reader can have a look at the chapter [18, §8.3] for a good introduction to "phylogenetic algebraic geometry".

The main observation that leads to the use of an algebraic geometry perspective is the following. The image of the map $\varphi_T$ defined in (2) "almost" fills an *algebraic variety* if we extend it to the complex field. An algebraic variety is the set of points where a collection of polynomials vanishes. The image of a polynomial map $\mathbb{C}^d \longrightarrow \mathbb{C}^{4^n}$ does not need to be an algebraic variety but it is always a *constructible set* (that is, in our case one needs to add to $Im\varphi_T$ some algebraic varieties of smaller dimension to obtain an algebraic variety). Let us call $V_T$ the smallest algebraic variety containing the image of $\varphi_T$ (i.e. the Zariski closure of $Im\varphi_T$). Then the polynomial equations that vanish on $Im\varphi_T$ are precisely those vanishing on $V_T$. Therefore, characterizing which distributions lie on the set $Im\varphi_T$ is equivalent to characterizing the intersection $V_T \cap \Delta$.

Summing up, the equations that define $V_T$ almost determine the set $Im\varphi_T$. Finding these equations is not an easy task. Actually, there are infinitely many equations that vanish on $V_T$ because the collection of equations that vanish on $V_T$ form an *ideal* $I_T$ in the ring of polynomials $\mathbb{R}[p_{\mathtt{AA\ldots A}}, p_{\mathtt{AA\ldots C}}, \ldots, p_{\mathtt{TT\ldots T}}]$. According to Hilbert's basis theorem, this ideal is finitely generated. Biologists Cavender, Felsenstein and Lake in [26] and [27] were the first to introduce the idea of using polynomials that vanish on any distribution on a tree $T$. They called these polynomials in the ideal of $V_T$ *phylogenetic invariants*. After a lot of efforts from algebraic geometers, collections of equations that define $V_T$ have been found for trees on any number of leaves evolving under the evolutionary models introduced in the previous section, see [20,21,28,29]. Those wanting to play with these phylogenetic invariants can have a look at webpage

https://www.coloradocollege.edu/aapps/ldg/small-trees/small-trees_0.html

which contains lists of generators for small trees, see [30].

The positive answer to the Questions on identifiability of the previous section ensure that any two of these varieties intersect properly (first question) and that the fibers of $\varphi_T$ are zero-dimensional (second question). Therefore, the dimension of $V_T$ (as algebraic variety, or as a manifold at the non-singular points) is equal to

the dimension of the parameter space. For the GM model, this dimension equals $d = 3+12|E(T)|$, so the codimension of $V_T$ is $4^n-3-12|E(T)|$, which is exponential on $n$ (the number of edges of an $n$-leaved tree is bounded above by $2n-3$). The number of polynomials in a minimal system of generators of the ideal of $V_T$ is too large to be used in practice: we need at least as many elements as the codimension to define a variety and we have just seen that this number is exponential on $n$. Moreover, it might be difficult to give a phylogenetic interpretation to some of these polynomials and to use them for phylogenetic reconstruction purposes. In Sect. 5 we explain other ways to use these ideals for phylogenetic reconstruction.

In what follows we give examples of phylogenetic invariants. For all the models and all trees, the following is a trivial phylogenetic invariant:

$$h : p_{\mathtt{AA\ldots A}} + p_{\mathtt{AA\ldots C}} + \cdots + p_{\mathtt{TT\ldots T}} - 1.$$

This polynomial vanishes on any point of $Im\varphi_T$ (and hence on the whole $V_T$) because probabilities must sum to one.

There are other phylogenetic invariants that depend only on the evolutionary model chosen but not on the tree structure. These are called *model invariants*. For example, for any tree on $n$-leaves evolving under the Jukes-Cantor model, the polynomial $p_{\mathtt{AA\ldots A}} - p_{\mathtt{CC\ldots C}}$ vanishes on any point $p^T \in Im\varphi_T$ (this can easily be seen using the symmetries of the transition matrices in JC69 model, and we can obtain other model invariants in the same way). Actually, if $p$ is a distribution that has arisen from a $G$-equivariant model evolving on a tree $T$, then

$$p_{g\mathtt{X}_1,\ldots,g\mathtt{X}_n} = p_{\mathtt{X}_1,\ldots,\mathtt{X}_n}$$

for any $g \in G$ (i.e. $p$ is $G$-invariant). Hence, $p_{g\mathtt{X}_1,\ldots,g\mathtt{X}_n} - p_{\mathtt{X}_1,\ldots,\mathtt{X}_n}$ are model invariants for any $g \in G$ and $\mathtt{X}_1,\ldots,\mathtt{X}_n \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^n$. These model invariants have been used in model selection in [31]: they have been implemented in a method that selects the $G$-equivariant model that best fits the data according to a statistical criterion.

The phylogenetic invariants that might be of interest in phylogenetic reconstruction are those that lie in $I_T$ but not in $I_{T'}$ for some other tree $T'$. These are called *topology invariants*.

For example, for the JC69 or K80 model on the tree $T_{12|34}$, Lake [27] found the following linear phylogenetic invariants:

$$\begin{aligned} H_1 :&\quad p_{xyxy} + p_{xyzw} - p_{xyzy} - p_{xyxw} \\ H_2 :&\quad p_{xyyx} + p_{xywz} - p_{xyyz} - p_{xywx} \end{aligned} \tag{3}$$

for any $x, y, z, w$ in $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$. It is not difficult to see that $H_1$ is not a phylogenetic invariant for $T_{13|24}$ and $H_2$ is not an invariant for $T_{14|23}$, so these polynomials are actually topology invariants for $T = T_{12|34}$. Lake used these two invariants to propose a method of phylogenetic reconstruction for quartet trees, without much success. It is not difficult to see why this method was not very successful: the variety $V_T$ is not a linear variety so using only linear invariants may not give the best results in phylogenetic reconstruction. In the next section we explain how to obtain other topology invariants of larger degrees.

## 3 Invariants from Flattenings

For the moment we consider trees on four leaves $\{1,2,3,4\}$. There are three possible (non-trivial) bipartitions of the set of leaves: 12|34, 13|24, 14|23. Let $p = (p_{\mathtt{AA...A}}, p_{\mathtt{AA...C}}, p_{\mathtt{AA...G}}, \ldots, p_{\mathtt{TT...T}})$ be a point in $\mathbb{R}^{4^4}$. Then we define the flattening of $p$ according to the bipartition 12|34 as the matrix

$$flat_{12|34}(p) = \begin{array}{c} \\ states \\ leaves \\ 1,2 \end{array} \overset{\begin{array}{c} states\ at\ \ leaves\ 3\ and\ 4 \end{array}}{\begin{pmatrix} p_{\mathtt{AAAA}} & p_{\mathtt{AAAC}} & p_{\mathtt{AAAG}} & \cdots & p_{\mathtt{AATT}} \\ p_{\mathtt{ACAA}} & p_{\mathtt{ACAC}} & p_{\mathtt{ACAG}} & \cdots & p_{\mathtt{ACTT}} \\ p_{\mathtt{AGAA}} & p_{\mathtt{AGAC}} & p_{\mathtt{AGAG}} & \cdots & p_{\mathtt{AGTT}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{\mathtt{TTAA}} & p_{\mathtt{TTAC}} & p_{\mathtt{TTAG}} & \cdots & p_{\mathtt{TTTT}} \end{pmatrix}}.$$

Note that this is a $16 \times 16$ matrix with rows labelled by the states $\mathtt{AA}, \mathtt{AC}, \ldots, \mathtt{TG}, \mathtt{TT}$ at leaves 1 and 2 and columns labelled by the states at leaves 3 and 4. We can define $flat_{13|24}(p)$ and $flat_{14|23}(p)$ analogously. Another way of interpreting these matrices is by using tensors: $p$ belongs to $\mathbb{R}^{4^4} \cong \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4$ and $flat_{12|34}(p)$ is the image of $p$ via the isomorphism from $(\mathbb{R}^4 \otimes \mathbb{R}^4) \otimes (\mathbb{R}^4 \otimes \mathbb{R}^4)$ to the set of linear maps $L((\mathbb{R}^4 \otimes \mathbb{R}^4)^*, \mathbb{R}^4 \otimes \mathbb{R}^4)$.

Using expression (1) one can see that, if $T = T_{12|34}$ and $p = \varphi_T(\pi; M^1, M^2, M^3, M^4, M^5)$ belongs to $Im\varphi_T$, then

$$flat_{12|34}(p) = (M^1 \otimes M^2)^t \, flat_{12|34}(q) \, M^3 \otimes M^4$$

where $M \otimes N$ denotes the Kronecker product of matrices and $q = \varphi_T(\pi; Id, Id, Id, Id, M^5)$. It can be easily seen that $flat_{12|34}(q)$ is a $16 \times 16$ matrix whose unique non-zero entries are labelled by $(\mathtt{XX}, \mathtt{YY})$. Hence, $flat_{12|34}(q)$ has rank $\leq 4$ and the same holds for $flat_{12|34}(p)$. This is the basis for the following result:

**Theorem 1.** *(Allman Rhodes, [32]) Let $p = \varphi_T(\pi; M^1, M^2, M^3, M^4, M^5)$ be a distribution arising from the GM model on $T = 12|34$. Then rank$(flat_{12|34}\, p) \leq 4$.*

*Moreover, rank$(flat_{13|24}\, p)$ and rank$(flat_{14|23}\, p)$ are equal to 16 if $M^i$ are invertible and $\pi$ is strictly positive.*

In other words, the $5 \times 5$ minors of $flat_{12|34}(p)$ are topology invariants for $T = T_{12|34}$. This result can be extended to larger trees by considering bipartitions $A|B$ of the set of leaves and flattening according to these bipartitions: the result would still refer to rank four matrices, as four is the number of states of the random variables we are considering at the nodes of the tree. Invariants from flattenings characterize whether a bipartition $A|B$ is present in the structure of the tree $T$ giving rise to the distribution $p$: if $T$ has an interior edge that partitions the set of leaves in $A|B$, then rank$(flat_{A|B}(p))$ is less than or equal to four.

An analogous version of this result has been proven for G-equivariant models in [33] by using techniques from representation theory. As a consequence, one can see that Lake's invariants in (3) appear from these rank conditions under

the JC69 model. Moreover, in the quoted paper it is proven that these rank conditions are enough to define a variety $V_{T_0}$ inside the union of varieties $\cup_{T\in\mathcal{T}} V_T$ (where $\mathcal{T}$ is the set of trees on the set of leaves $\{1,\ldots,n\}$):

**Theorem 2.** *(Casanellas, Fernández-Sánchez 2011) Consider any $G$-equivariant model. For each tree topology $T \in \mathcal{T}$, there exists a dense open subset $U_T \subseteq V(T)$ such that if a point $p$ belongs to $\bigcup_{T\in\mathcal{T}} U_T$, then $p$ belongs to $V_{T_0}$ if and only if the rank of all $flat_{A|B}(p)$ is less than or equal to four for all bipartitions induced from interior edges of $T_0$.*

Roughly speaking, if one assumes that $p$ has arisen as a Markov process on some phylogenetic tree (with generic parameters), considering invariants from flattenings is enough to detect the tree topology. As assuming that $p$ has arisen on some phylogenetic tree with generic parameters (i.e. $p \in \bigcup_{T\in\mathcal{T}} U_T$ ) is what all common phylogenetic reconstruction methods do, this theorem says that considering rank conditions from flattenings is sufficient for tree topology reconstruction via algebraic geometry.

> **Recall**

Assuming that data comes from a phylogenetic tree, it is enough to consider rank conditions from flattenings to recover the tree (see Theorem 2).

In Sect. 5 below, we shall see an alternative way to use these rank conditions in practice.

## 4 Semi-algebraic Constraints

As the reader may have realized, when passing from the Markov process on a tree $T$ to the map $\varphi_T$ in (2), we are eluding the fact that the parameters are probabilities. Let us call $\varphi_T^+$ the restriction of $\varphi_T$ to *stochastic parameters* (non-negative parameters that sum to one) and let $V_T^+ = Im\varphi_T^+$. The map $\varphi_T^+$ defines a *semi-algebraic variety*, that is, it can be described by polynomial equations and polynomial inequalities. In [34] we examined whether it is relevant to take this into account and we concluded that there are biologically realistic situations where considering the whole algebraic variety (and not only this part $V_T^+$) can be misleading. These situations are actually the typical scenarios where most phylogenetic reconstruction methods fail.

In [15], Allman, Rhodes and Taylor have characterized those distributions that lie in $V_T^+$. To this end, given a tensor $p \in \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4$ they consider some transformations $\tilde{p}$ of $p$ obtained by multiplying trivial flattenings of $p$ (considering bipartitions of one leaf against the others) by matrices coming from the marginalization of $p$ over two indices. If $p = \varphi_T(\pi; M_1, M_2, M_3, M_4, M_5)$ is

in $Im\varphi_T$, $T = T_{12|34}$, these transformations produce a point $\tilde{p}$ also in $Im\varphi_T$ but with different transition matrices: $\tilde{p} = \varphi_T(\pi; M, M, N, N, M_5)$ where $M = M_1$ or $M_2$ and $N = M_3$ or $M_4$. We state here their result in the case of trees with four leaves.

**Theorem 3.** *(Allman, Rhodes and Taylor, 2014) A point $p \in \mathbb{R}^{4^4}$ arises from a Markov process on the phylogenetic tree $T_{12|34}$ with stochastic parameters if and only if:*

*(i)* $flat_{12|34}(p)$ *has rank* $\leq 4$,
*(ii) the marginalizations* $p_{+\ldots}$ *and* $p_{\ldots+}$ *arise from stochastic parameters, and*
*(iii)* $flat_{13|24}(\tilde{p})$ *is positive semi-definite, where* $\tilde{p}$ *is any of the transformations mentioned above.*

In the next section we will explain how to use this result in phylogenetic reconstruction.

## 5 Phylogenetic Reconstruction

We come back to the problem of phylogenetic reconstruction. Our input data are DNA sequences of length $N$ given in the form of an *alignment*. An alignment of $n$ DNA sequences is an $n \times N$ array whose rows correspond to the DNA sequences and whose columns represent nucleotides that have evolved from the same nucleotide at the common ancestor of the sequences. For example, in Fig. 4 we have an alignment of four DNA sequences; the first nucleotide in each sequence corresponds to a certain nucleotide at the common ancestor of these sequences (probably an `A` as well).

$s_1$ : AACTTCGAGGCTTACC
$s_2$ : AAGGTCGATGCTCACC
$s_3$ : AACGTCTATGCTCACC
$s_4$ : GACGCCGATGCTCATC

**Fig. 4.** An alignment of four DNA sequences

As we assumed that all positions in a DNA sequence evolve independently and in the same way, an alignment can be thought of as N independent samples from a multinomial distribution $p = (p_{\mathtt{AA\ldots A}}, p_{\mathtt{AA\ldots C}}, \ldots, p_{\mathtt{TT\ldots T}})$. This distribution can be estimated from the relative frequencies $f = (f_{\mathtt{AA\ldots A}}, \ldots, f_{\mathtt{TT\ldots T}})$ of each column $\mathtt{X}_1, \ldots \mathtt{X}_n$ in the alignment, and $f \sim p$ when $N$ tends to infinity. If these sequences had evolved according to a Markov process on a tree $T$ (for certain parameters), then $p$ would belong to $Im\varphi_T^+$ and $f$ would be close to $V_T^+$.

The aim is to use the results in the previous sections that characterized distributions in $V_T$ and $V_T^+$ to design methods that reconstruct $T$ from the vector of relative frequencies $f$. A phylogenetic reconstruction method is called *statistically consistent* if it outputs $T$ with probability one when the alignment has

been generated by $p \in Im\varphi_T^+$ and we let the length of the alignment tend to infinite. This is the least we should require to a phylogenetic reconstruction method. Designing a reconstruction method that is statistically consistent is not so difficult, the difficult part is to make it reach convergence fast enough because sequences at our disposal are often not very long.

One of the most common phylogenetic reconstruction methods is *maximum likelihood estimation*. Given an alignment $D$ and an evolutionary model, the method seeks to obtain the tree topology $T_0$ and the substitution parameters $\hat{\theta} = (\pi; \{M^e\}_e)$ which maximize $Prob(D|T, \theta)$ among all possible phylogenetic trees $T$ and substitution parameters $\theta$. To this end, first the maximum likelihood estimate of the substitution parameters is obtained separately for each tree topology $T$ (using some of the available optimization methods) and then one chooses the tree topology and the parameter estimates which maximize the likelihood among all tree topologies.

This method has a clear drawback: the number of (isomorphism classes of) trivalent phylogenetic trees on $n$ leaves is $(2n-5)!!$, which grows exponentially in $n$, so that it becomes unfeasible to do an exhaustive search through all tree topologies for more than 20 leaves (even with nowadays computational capacity). The vast majority of phylogenetic reconstruction software use some branch and bound algorithm and only cover a small part of the tree space. On the other hand, numerical optimization methods do not guarantee a global maximum in general and, moreover, it is known that there are multiple local maximum for biologically relevant parameters.

By far, the most used phylogenetic reconstruction method is Neighbor-Joining [35]. This a *distance-based method*, that is, all the information from an alignment on a set of species $\{1, \ldots, n\}$ is condensed into a *dissimilarity function* $d : \{1, \ldots, n\} \times \{1, \ldots, n\} \longrightarrow \mathbb{R}_{\geq 0}$ (symmetric and with zero diagonal entries). This dissimilarity function is intended to approximate the evolutionary distance between pairs of species or, in other words, it should account for the amount of elapsed substitutions between both species. Obviously, not all substitutions that have occurred during evolution can be observed in the contemporary species sequences (for instance, there may be an `A` mutating to `T` and finally coming back to `A` in the nowadays species) and the dissimilarity function has to take this into account. For example, the *Jukes-Cantor distance* between two DNA sequences defined as $-\frac{3}{4}\ln(1-\frac{4}{3}s)$, where $s$ is the fraction of nucleotides that differ in both sequences. It approximates the amount of substitutions (observed and unobserved) between the species if these have evolved under the Jukes-Cantor model.

Given a dissimilarity function $d$, the first step in the Neighbor-Joining algorithm chooses two species $x$ and $y$ minimizing function $D_{x,y} = d(x, y) - \frac{1}{n-2}\sum_z (d(x, z) + d(y, z))$, seeking to minimize their distance but maximize the average distance to the other species. These two species are joined on a cherry (that is, two leaves joined by two edges and an interior node) and the interior node is treated as a new species substituting the former $x$ and $y$. In this way

the number of species is decreased at each step and the function $D$ is redefined accordingly.

This algorithm produces the correct phylogenetic tree if the input dissimilarities can be realized as sums of lengths on the path between leaves on a phylogentic tree. However, when dealing with biological sequences, their estimated distances do not correspond to the branch lengths of any particular tree, and the tree constructed by Neighbor-joining algorithm may not have a realistic biological interpretation. In spite of this, it is one of the most widely used method and, as there is no need to search through the whole space of phylogenetic trees, it is used to produce phylogenetic trees for large number of species.

In the next section we explain of to use an algebraic geometry approach to produce alternative methods of phylogenetic reconstruction.

### 5.1 Reconstruction of Quartets

We start by reviewing algebraic reconstruction methods for quartet trees and in the next subsection we explain how to get to larger trees.

One of the most simple ways to use rank conditions (Theorem 1) in phylogenetic reconstruction is via the singular value decomposition of the flattening matrix [36]. A *singular value decomposition* of an $m \times n$ matrix $M$ is a decomposition $M = UDV^t$ where $U, V$ are orthogonal matrices and $D$ is an $m \times n$ matrix with off-diagonal entries equal to zero and ordered diagonal entries $D_{1,1} = \sigma_1 \geq \sigma_2 \geq \cdots \geq D_{r,r} = \sigma_r > 0$, $D_{i,i} = 0$ for $i > r = rank M$. The elements $\sigma_1, \ldots, \sigma_r$ are called singular values and are uniquely determined by $M$. The Eckart-Young theorem [37] states that the distance (in Frobenius norm) from a matrix $M$ to the set of matrices of rank $\leq k$ is given by

$$\delta_k(M) = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2}.$$

According to Theorem 1, if $T = 12|34$ and $p \in Im\varphi_T$ has been generated from generic parameters, we have $\delta_4(flat_{12|34}(p)) = 0$ and $\delta_4(flat_{13|24}(p)) > 0$, $\delta_4(flat_{14|23}(p)) > 0$. This idea was explored in [38] to propose a method of phylogenetic reconstruction: given the vector $f$ of relative frequencies obtained from an alignment of DNA sequences, compute the triplet of scores

$$s_{12|34}(f) = \delta_4(flat_{12|34}(f)), \quad s_{13|24}(f) = \delta_4(flat_{13|24}(f)),$$

$$s_{14|23}(f) = \delta_4(flat_{14|23}(f))$$

and output the tree $T_{A|B}$ if $s_{A|B}(f)$ is the smallest.

In [39] this method was further exploited and a modification was proposed to make it more efficient and more robust. Despite of this improvement, methods based solely on rank conditions seem to need large alignments to outperform other reconstruction methods. That is, the method presented is statistically consistent but it converges slowly to the correct tree.

As argued in [34], methods based only on algebraic conditions may have problems in reconstructing the correct tree from small samples because they do

not take into account that parameters generating the alignment must be positive. However, incorporating semi-algebraic conditions does not seem an easy task. Out of the three conditions stated in Theorem 3, we are mostly interested in the (i) and (iii), as the second one is not directly related to the tree structure. The first is the rank conditions we discussed above. In order to incorporate the third condition we borrow the following result from linear algebra:

**Theorem 4.** *(Higham [40]) Let $M$ be an $n \times n$ matrix. Consider its closest symmetric matrix $S = \frac{M+M^t}{2}$ and the polar decomposition of $S$, $S = UH$, where $U$ is orthogonal and $H$ is a positive semidefinite matrix. Then $psd(M) = \frac{S+H}{2}$ is the closest positive semidefinite matrix to $M$ (in Frobenius norm).*

This theorem may allow us to compute how far is $flat_{13|24}(\tilde{p})$ from the set of positive semidefinite matrices. However, the goal is to combine conditions (i) and (iii) in Theorem 3 into a single score. To this aim, we proved in [41] that the rank of $psd(M)$ is always smaller than or equal to the rank of $M$. With all this in mind, in [16] we came out with the following method: given the vector of relative frequencies $f$ of an alignment, compute all transformations of the vector mentioned in Theorem 3 (see [16] for details) and for each transformation $\tilde{f}$ compute

$$s_{12|34}(\tilde{f}) := \frac{\min\left\{\delta_4\left(psd\left(flat_{13|24}\left(\tilde{f}\right)\right)\right), \delta_4\left(psd\left(flat_{14|23}\left(\tilde{f}\right)\right)\right)\right\}}{\delta_4\left(psd\left(flat_{12|34}\left(\tilde{f}\right)\right)\right)}. \quad (4)$$

Then we define the score $s_{12|34}(f)$ of the alignment as the average of these scores and define $s_{13|24}(f)$, $s_{13|24}(f)$ accordingly. Note that for a distribution $p \in Im\varphi_T^+$ with $T = T_{12|34}$, we have $s_{12|34}(p) = \infty$. Indeed, according to Theorem 3(iii), $flat_{13|24}(\tilde{p})$ is positive semidefinite, so

$$rank\left(psd\left(flat_{13|24}(\tilde{p})\right)\right) = rank\left(flat_{13|24}(\tilde{p})\right).$$

As any transformation $\tilde{p}$ of $p$ belongs to $V_T$, this rank is greater than 4 (for generic parameters). This proves that the numerator is different than zero. On the other hand, as mentioned above, the rank of $psd\left(flat_{12|34}(\tilde{p})\right)$ is bounded above by the rank of $flat_{12|34}(\tilde{p})$, which is smaller than or equal to four (because $\tilde{p} \in V_T$). Therefore, the denominator of 4 is zero and $s_{12|34}(p) = \infty$.

With the same arguments we can prove that the other two scores $s_{13|24}(p)$ and $s_{14|23}(p)$ are zero. Thus, we can consider a method that outputs the tree $T_{A|B}$ for which $s_{A|B}(f)$ is largest. This is called the SAQ (for Semi-Algebraic Quartet) method in [16]. The previous argument proves that SAQ is a statistically consistent method. Moreover, it is much more efficient than previous algebraic methods and outperforms maximum-likelihood and neighbor-joining for alignments of length at least 500 generated under the GM model (whereas a method that only considers rank conditions needs at least 10000 sites to beat other methods).

### 5.2 From Quartets to Larger Trees

So far we have just seen algebraic methods that reconstruct only trees with four leaves (quartets). We are interested now in reconstructing phylogenetic trees with $n$ leaves. Some methods, like Neighbor-Joining introduced above, build directly a tree with $n$ leaves from an alignment of DNA sequences. But other methods built an $n$-leaved tree out of smaller pieces, namely quartet trees. These are called *quartet-based methods*.

We briefly describe one of these methods, Weight Optimization (WO for short), see [42]. WO needs as input a triplet of weights for each four species. That is, starting with an alignment of $n$ DNA sequences, we need to consider all 4-tuples $i, j, k, l$ of the sequences and use a reconstruction method to weight the three possible quartet trees: $w_{ijkl} = (w_{ij|kl}, w_{ik|jl}, w_{il|jk})$, each weight expressing the reliability of the corresponding quartet tree. For example, we can use SAQ to produce input weights for WO by defining:

$$w_{ijkl} = \left( \frac{s_{T_{ij|kl}}(f)}{s(f)}, \frac{s_{T_{ik|jl}}(f)}{s(f)}, \frac{s_{T_{il|jk}}(f)}{s(f)} \right)$$

where $s(f) = s_{T_{ij|kl}}(f) + s_{T_{ik|jl}}(f) + s_{T_{il|jk}}(f)$. In this way, weights are normalized between 0 and 1.

Then WO randomly chooses a starting 4-tuple $i, j, k, l$ and dynamically defines the species addition order seeking to maximize the total sum of weights at each step: the added species at each step is selected and placed at the edge that gives the largest possible increase of weight. WO is known to reconstruct the correct tree if the quartets are correctly weighted. We can use SAQ to weight quartet trees but we could also use a method based only on rank conditions or on maximum likelihood strategies. As WO is highly dependant on the initial quartet, usually the method is run for several trials, say 100, so that 100 $n$-leaved trees are produced. These trees might not coincide and might be incompatible but a *consensus tree* can be built out of them (following a chosen criterion, such as the majority rule consensus tree, see [43]).

We ran WO with SAQ on the DNA sequences from Ratites used in [2] for 100 trials. The majority rule consensus tree obtained is the tree depicted in Fig. 1 (see [17]). Therefore, our methods support the hypothesis that Tinamous evolved separately from the rest of Ratites (see the discussion at the legend of the figure).

## 6 Discussion

In this report we have presented probabilistic models of nucleotide substitution that allow approaching phylogenetic reconstruction from an algebraic point of view. We have introduced some of the main techniques that are used in the development of algebraic and semi-algebraic tools in phylogenetics and we have seen how these tools can be used in the reconstruction of the tree topology of a set of DNA sequences.

Still, there are problems for which algebraic phylogenetics has not been fully developed yet. For instance, it would be interesting to provide tree reconstruction methods for $G$-equivariant models, not only for the general Markov model. This would be specially interesting in the case of protein data, as the general Markov model of amino acid substitution seems too general to treat these data: it will involve 380 parameters per edge whereas current models used by biologists only involve one. Developing algebraic tools for models in between these two would certainly be of interest.

Also, we have assumed that sites in the sequences were identically distributed. Considering mixtures of distributions is a way of relaxing these assumptions. Mixtures of distributions are just convex combinations of distributions, and therefore also allow an algebraic approach. Further studying mixtures for different evolutionary models can lead to more flexibility in the models where algebraic tools can be applied. We need to mention that the tools presented here work for gene trees, and one has to combine them with a coalescent model in order to infer species trees.

On the other hand, we have mentioned how to reconstruct trees from quartets inferred by algebraic tools. It would be desirable to extend algebraic techniques to develop methods of reconstruction of large trees directly, not passing through quartets.

**Acknowledgments.** The author would like to thank her coauthors Jesús Fernández-Sánchez and Marina Garrote-López for all the joint work described here. The author has been partially supported by PID2019-103849GB-I00 and CEX2020-001084-M of AEI, Government of Spain.

## References

1. Phillips, M.J., Gibb, G.C., Crimp, E.A., Penny, D.: Tinamous and Moa flock together: mitochondrial genome sequence analysis reveals independent losses of flight among ratites. Syst. Biol. **59**(1), 90–107 (2009)
2. Vera-Ruiz, V.A., Robinson, J., Jermiin, L.S.: A likelihood-ratio test for lumpability of phylogenetic data: is the Markovian property of an evolutionary process retained in recoded DNA? Syst. Biol. **71**(3), 660–675 (2021)
3. Jermiin, L.S., Catullo, R.A., Holland, B.R.: A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. NAR Genom. Bioinform. **2**(2) (2020)
4. Makarenkov, V., Mazoure, B., Rabusseau, G., Legendre, P.: Horizontal gene transfer and recombination analysis of SARS-CoV-2 genes helps discover its close relatives and shed light on its origin. BMC Ecol. Evol. **21**, 2730–7182 (2021)
5. Pistone, G., Riccomagno, E., Wynn, H.P.: Algebraic Statistics: Computational Commutative Algebra in Statistics. Chapman and Hall/CRC, Boca Raton (2000)
6. Pachter, L., Sturmfels, B.: Parametric inference for biological sequence analysis. Proc. Natl. Acad. Sci. **101**(46), 16138–16143 (2004)
7. Pachter, L., Sturmfels, B.: Tropical geometry of statistical models. Proc. Natl. Acad. Sci. **101**(46), 16132–16137 (2004)
8. Allman, E.S., Ane, C., Rhodes, J.A.: Identifiability of the GTR+\Gamma model of molecular evolution. arXiv e-prints, 709, September 2007

9. Sumner, J.G., Jarvis, P.D., Fernández-Sánchez, J., Kaine, B.T., Woodhams, M.D., Holland, B.R.: Is the general time-reversible model bad for molecular phylogenetics? Syst. Biol. **61**(6), 1069–1074 (2012)
10. Casanellas, M., Fernández-Sánchez, J., Roca-Lacostena, J.: The embedding problem for Markov matrices. Publicacions Matemàtiques (2022, to appear)
11. Casanellas, M., Fernández-Sánchez, J., Michałek, M.: Local equations for equivariant evolutionary models. Adv. Math. **315**, 285–323 (2017)
12. Friedland, S., Gross, E.: A proof of the set-theoretic version of the salmon conjecture. J. Algebra **356**(1), 374–379 (2012)
13. Michałek, M., Ventura, E.: Phylogenetic complexity of the kimura 3-parameter model. Adv. Math. **343**, 640–680 (2019)
14. Casanellas, M., Garrote-López, M., Zwiernik, P.: Identifiability in robust estimation of tree structured models. Bernoulli (2022, to appear)
15. Allman, E.S., Rhodes, J.A., Taylor, A.: A semialgebraic description of the general Markov model on phylogenetic trees. SIAM J. Discret. Math. **28**(2), 736–755 (2014)
16. Casanellas, M., Fernández-Sánchez, J., Garrote-López, M.: SAQ: semi-algebraic quartet reconstruction method. IEEE/ACM Trans. Comput. Biol. Bioinf. **18**(6), 2855–2861 (2021)
17. Casanellas, M., Fernández-Sánchez, J., Garrote-López, M.: Designing weights for quartet-based methods when data is heterogeneous across lineages. Bull. Math. Biol. **85**, 68 (2023)
18. Steel, M.A.: Phylogeny: Discrete and Random Processes in Evolution. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2016)
19. Barry, D., Hartigan, J.A.: Asynchronous distance between homologous DNA sequences. Biometrics **43**(2), 261 (1987)
20. Draisma, J., Kuttler, J.: On the ideals of equivariant tree models. Math. Ann. **344**(3), 619–644 (2009)
21. Casanellas, M., Sullivant, S.: The strand symmetric model. In: Pachter, L., Sturmfels, B. (eds.) Algebraic Statistics for Computational Biology, chap. 16, pp. 305–321. Cambridge University Press, Cambridge (2005)
22. Kimura, M.: Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. **78**(1), 454–458 (1981)
23. Kimura, M.: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**(2), 111–120 (1980)
24. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. Mamm. Protein Metab. **3**, 21–132 (1969)
25. Chang, J.T.: Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. **137**(1), 51–73 (1996)
26. Cavender, J.A., Felsenstein, J.: Invariants of phylogenies in a simple case with discrete states. J. Classif. **4**(1), 57–71 (1987)
27. Lake, J.A.: A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol. Biol. Evol. **4**, 167–191 (1987)
28. Allman, E.S., Rhodes, J.A.: Phylogenetic ideals and varieties for the general Markov model. Adv. Appl. Math. **40**(2), 127–148 (2008)
29. Sturmfels, B., Sullivant, S.: Toric ideals of phylogenetic invariants. J. Comput. Biol. **12**(2), 204–228 (2005)
30. Casanellas, M., Garcia, L.D., Sullivant, S.: Catalog of small trees. In: Pachter, L., Sturmfels, B. (eds.) Algebraic Statistics for Computational Biology, chap. 15, pp. 305–321. Cambridge University Press, Cambridge (2005)

31. Kedzierska, A.M., Drton, M., Guigó, R., Casanellas, M.: SPIn: model selection for phylogenetic mixtures via linear invariants. Mol. Biol. Evol. **29**(3), 929–937 (2012)
32. Allman, E.S., Rhodes, J.A.: Phylogenetic invariants for the general Markov model of sequence mutation. Math. Biosci. **186**(2), 113–144 (2003)
33. Casanellas, M., Fernández-Sánchez, J.: Relevant phylogenetic invariants of evolutionary models. Journal de Mathématiques Pures et Appliquées **96**(3), 207–229 (2011)
34. Casanellas, M., Fernández-Sánchez, J., Garrote-López, M.: Distance to the stochastic part of phylogenetic varieties. J. Symb. Comput. **104**, 653–682 (2021)
35. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**, 406–425 (1987)
36. Meyer, C.D.: Matrix analysis and applied linear algebra. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2000)
37. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika **1**(3), 211–218 (1936)
38. Eriksson, N.: Tree construction using singular value decomposition. In: Pachter, L., Sturmfels, B. (eds.) Algebraic Statistics for Computational Biology, chap. 19, pp. 347–358. Cambridge University Press, Cambridge (2005)
39. Fernández-Sánchez, J., Casanellas, M.: Invariant versus classical approach when evolution is heterogeneous across sites and lineages. Syst. Biol. **65**, 280–291 (2016)
40. Higham, N.J.: Matrix nearness problems and applications. In: Applications of Matrix Theory, vol. 22 (1989)
41. Casanellas, M., Fernández-Sánchez, J., Garrote-López, M.: The inertia of the symmetric approximation for low-rank matrices. Linear Multilinear Algebra **66**(11), 2349–2353 (2018)
42. Ranwez, V., Gascuel, O.: Quartet-based phylogenetic inference: improvements and limits. Mol. Biol. Evol. **18**(6), 1103–1116 (2001)
43. Barrett, M., Donoghue, M.J., Sober, E.: Against consensus. Syst. Zool. **40**(4), 486–493 (1991)