# ALGEBRAIC TECHNIQUES FOR PHYLOGENETIC RECONSTRUCTION

MARTA CASANELLAS, JESÚS FERNÁNDEZ-SÁNCHEZ, AND MARINA GARROTE-LÓPEZ

Abstract. In the last years there has been an increase in the use of algebraic tools in phylogenetic reconstruction. As a Markov process of substitution of nucleotides along a phylogenetic tree can be viewed as a polynomial map, it is natural to use tools from algebraic geometry and computational algebra in this setting.

We will explain how these tools are used in phylogenetics and we will introduce the use of semi-algebraic constraints, which mirror the probabilistic nature of the parameters used. We will show advantages and disadvantages of the use of these algebraic methods using real and simulated data.

## Introduction

Phylogenetics studies the evolutionary history of living species. Species evolution can be represented in a *phylogenetic tree* where leaves stand for current species and interior nodes represent their ancestral species. Phylogenetic inference is done via genomic sequences and has an impact beyond scientific curiosity: it is used to identify the origin of new pathogens (as it has been the case for SARS-Cov2 virus), it helps designing biodiversity conservation policies, and can be useful on the traceability of cancer cells, among other examples.

Markov processes are the natural way of modeling the substitution of nucleotides (or other characters) along a phylogenetic tree $T$. These processes can be understood as polynomial maps $\varphi_T$ between two affine spaces, whose images contain the set of distributions on the leaves of the tree that arise from such Markov processes. By studying these images $Im\varphi_T$ from the point of view of algebraic varieties, one can provide tools that allow to distinguish between distributions that have arisen on different trees (i.e. tools to decide which tree is most plausible to have generated the given data).

## 1. Algebraic constraints

Allman and Rhodes proved that, if $p$ is a distribution that has arisen on a tree $T$ and $e$ is an interior edge of $T$ that splits the leaves of $T$ in $A|B$, then *flattening* $p$ according to the distribution $A|B$ gives rise to a matrix of rank four at most (see [1]). On the contrary, flattening $p$ according to a bipartition not consistent with any edge of $T$, gives rise to a rank

The first author has been partially supported by Spanish Government Agencia Estatal de Investigación, PID2019-103849GB-I00 AEI and CEX2020-001084-M AEI..

The second author has been partially supported by Spanish Government Agencia Estatal de Investigación, PID2019-103849GB-I00 AEI and CEX2020-001084-M AEI..

The third author has been partially supported by Spanish Government Agencia Estatal de Investigación, PID2019-103849GB-I00 AEI..

The talk at the 8IMM 2022 has been given by the first author.

16 matrix (in general). Thus, looking at the rank of these flattening matrices gives a clue for reconstructing the tree structure. The original result of Allman and Rhodes worked for the general Markov model of nucleotide substitution and was extended to other models in [3].

These tools have been used to develop a few phylogenetic reconstruction methods, for example Erik+2 ([6]) which considers a normalization of these flattening matrices. This method is restricted to four-leaved trees, but it provides weights that can be used as input of *quartet-based methods* (methods that build large phylogenetic trees by using as key ingredients the reconstruction of four-leaved trees).

## 2. Semi-algebraic constraints

On a related work, Allman, Rhodes and Taylor [2] studied the constraints on $Im\varphi_T$ that are a consequence of the probabilistic nature of the parameters in the domain. These parameters must be non-negative, which implies that their image through $\varphi_T$ lies in a semi-algebraic variety. The semi-algebraic constraints can be easily be translated on flattening matrices with linear algebra techniques. They have been used to design in a new phylogenetic reconstruction method SAQ, see [4], which considers both the rank conditions mentioned above and these semi-algebraic constraints. Although this method is targeted for four-leaved trees, it also provides weights to be used as input of quartet-based methods and we have recently implemented it in [5].

We shall see how these methods work and, with real and simulated data, what is the improvement of SAQ over Erik+2. Moreover, we shall discuss the advantages of considering algebraic tools against traditional tools for phylogenetic reconstruction.

## References

[1] E. S. Allman and J. A. Rhodes. *Phylogenetic invariants.* In Gascuel, O. and Steel, M. A., editors, *Reconstructing Evolution.* Oxford University Press, 2007.

[2] Allman, E. S., Rhodes, J. A., and Taylor, A. (2012). A semialgebraic description of the general Markov model on phylogenetic trees. *SIAM Journal on Discrete Mathematics*, 28.

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

[4] M. Casanellas, J. Fernández-Sánchez, and M. Garrote-López. *SAQ: semi-algebraic quartet reconstruction method.* IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021), 1–28.

[5] M. Casanellas, J. Fernández-Sánchez, M. Garrote-López, and M. Sabaté-Vidales. *Designing weights for quartet-based methods when data is heterogeneous across lineages.* https://arxiv.org/abs/2202.13365.

[6] Fernández-Sánchez, J. and Casanellas, M. (2016). Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. *Systematic Biology*, 65(2):280–291.

Marta Casanellas; Universitat Politècnica de Catalunya
*Email address*: marta.casanellas@upc.edu

Jesús Fernández-Sánchez; Universitat Politècnica de Catalunya
*Email address*: jesus.fernandez.sanchez@upc.edu

Marina Garrote-López; University of British Columbia
*Email address*: mgarrote@math.ubc.ca