The positive part of phylogenetic varieties

Marta Casanellas, Jesús Fernández-Sánchez, **Marina Garrote-López**

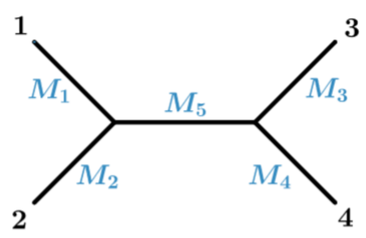
Universitat Politècnica de Catalunya - Barcelona Graduate School of Mathematics

Introduction

Modelling the substitution of nucleotides along a phylogenetic tree is usually done by a hidden Markov process. This allows to define a distribution of characters at the leaves of the trees and one might be able to obtain polynomial relationships among the probabilities of different characters. The study of these polynomials and the geometry of the algebraic varieties that define can be used to reconstruct phylogenetic trees. However, not all points in these algebraic varieties have biological sense. We explore the extent to which adding semialgebraic conditions arising from the restriction to parameters with statistical meaning can improve existing methods of phylogenetic reconstruction. Here we present our results of the computation of the distance of data points to the Jukes Cantor algebraic varieties and to the stochastic part of theses varieties.

Phylogenetic trees

Let T be 4-leaf a phylogenetic tree. There are three possible tree topologies, 12|34, 13|24, and 14|23, according to the shape of the tree taking into account the names of the species at the leaves. Suppose $T = T_{12|34}$ and the evolutionary process on that tree follows an nucleotide substitution model $\mathcal{M}$ associate a random variable x_i taking values on $\Sigma := \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ at each node, and consider as parameters distribution π at the root, and transition matrix M_e at each edge of T.



The joint distribution is a vector $P \in \mathbb{R}^{4^4}$ whose entries are the observed joint probabilities at the leaves:

$$p_{xyzt} = Prob(x_1 = x \ \& \ x_2 = y \ \& \ x_3 = z \ \& \ x_4 = t),$$

and can be estimated as relative frequencies from an alignment.

Phylogenetic Varieties

We denote by ψ_T the parametrization map:

$$\psi_T : S \subset [0,1]^\ell \to \mathbb{R}^{4^4}$$
$$\{\pi, \{M_e\}_{e \in E(T)}\} \mapsto P = (p_{\mathtt{AAAA}}, p_{\mathtt{AAAC}}, \dots, p_{\mathtt{TTTG}}, p_{\mathtt{TTTT}})$$

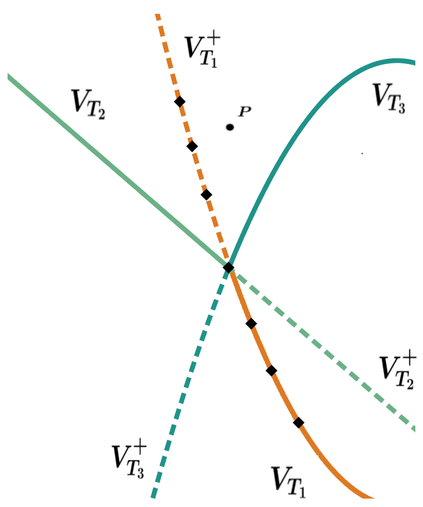
which maps each set of parameters of the model $\{\pi, \{M_e\}_{e \in E(T)}\}$ to the joint distribution of characters at the leaves of T.

Define the phylogenetic variety associated with T as the smallest algebraic variety containing $\psi_T(\mathbb{R}^l)$,

$$\mathcal{V}_T = \overline{\psi_T(\mathbb{R}^l)}.$$

And the stochastic phylogenetic variety as the set corresponding to all distributions arising from stochastic parameters:

$\mathcal{V}_T^+ = \{P \in \mathcal{V}_T \mid P = \psi_T(s) \text{ and } s \in S\}$.



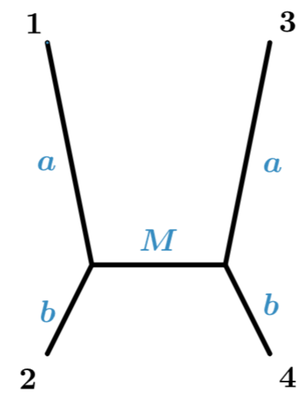
Simulations with the Jukes Cantor Model

The Jukes-Cantor (*JC*) model assumes the uniform distribution at the root $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and transition matrices:

$$M = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}, \text{ where } a + 3b = 1,$$

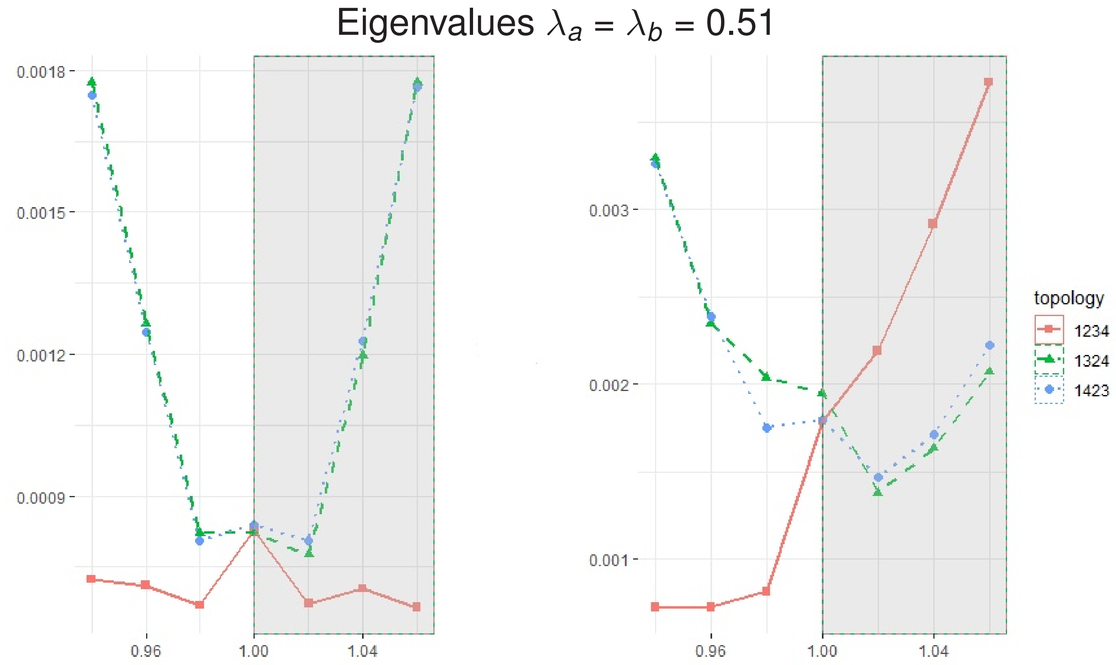
that can be parametrized by the eigenvalue of M, $\lambda(M)$ different from 1.

We consider a 4-leaf tree $T_{12|34}$ with *JC* matrices. Suppose λ_a and λ_b are the eigenvalues of matrices at the exterior edges and M is a *JC* matrix at the interior edge, with eigenvalue λ_m that takes values in the interval $[0.94, 1.06]$

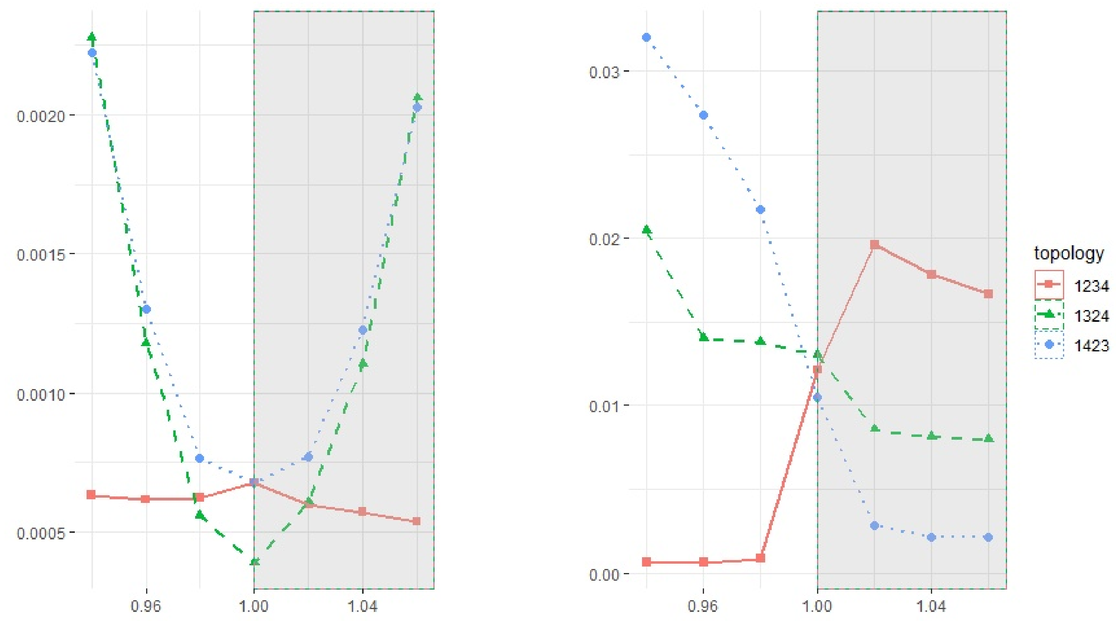


Distance to phylogenetic varieties

Eigenvalues $\lambda_a = \lambda_b = 0.51$



Eigenvalues $\lambda_a = 0.37$ and $\lambda_b = 0.87$



On the left: distance to the phylogenetic varieties $\mathcal{V}_T$. On the right: distance to the stochastic part of the varieties $\mathcal{V}_T^+$. The x-axis represents the values of λ_m.

References

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

Casanellas, M., and Fernandez-Sanchez, J. *Geometry of the Kimura 3-parameter model.* Advances in Applied Mathematics 41, 2008.

Draisma, J., Horobet, E., Ottaviani, G., Sturmfels, B., and Thomas, R. *The euclidean distance degree of an algebraic variety*. Foundations of Computational Mathematics, 2015.

Kosta, D., and Kubjas, K. *Maximum likelihood estimation of symmetric group-based models via numerical algebraic geometry*. Bulletin of Mathematical Biology 81, 2, 2019.

Contact: `marina.garrote@upc.edu`