

Homogeneous Phylogenetic Models: Invariants and Parametric Inference

Nicholas Eriksson

Department of Mathematics, UC Berkeley
eriksson@math.berkeley.edu



Overview

Homogeneous Phylogenetic Models:

- Rooted phylogenetic tree with n nodes.
- Each node is a binary, observed random variable Y_i .
- Every edge has the same transition matrix $A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$
- The joint probabilities are $p_{\sigma_1 \sigma_2 \dots \sigma_n} = P(\mathbf{Y} = \sigma)$. In terms of the parameters, the probability of observing σ is the product of the parameters a_{ij} that correspond to the transitions between the σ_i on the edges of the tree.

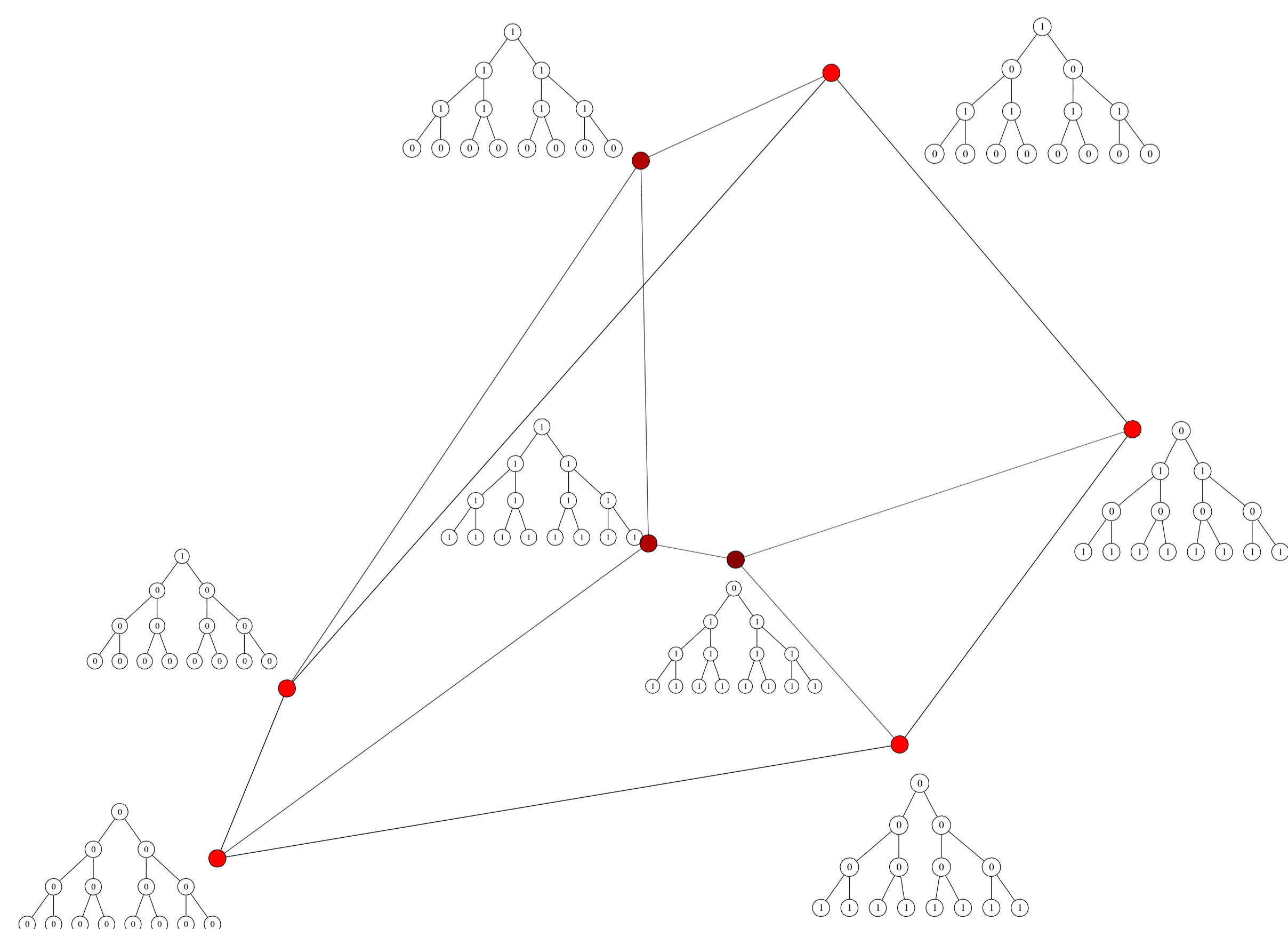
Questions:

We are interested in two questions from [6] that are of fundamental importance to the use of statistical models in biology.

1. Given observations $\sigma = (\sigma_1, \dots, \sigma_n)$, describe the set of parameters a_{ij} such that p_σ is maximal among the coordinates of p .
2. Which (parameter independent) relations on the probabilities $p(\mathbf{Y} = \sigma)$ does the model imply?

Example

The figure is a convex polytope with 8 vertices. Each vertex corresponds to a labeling of the tree that is of maximal probability for some choice of parameters. Next to each vertex is the corresponding tree.



THEOREM: (ERIKSSON, [2]) *If T is any binary tree with more than 3 nodes, in which every leaf is an odd distance from the root, then there are exactly eight labelings of the tree which are the labelings of maximal probability for some choice of parameters.*

Parametric Inference

Question:

Given any labeling $\mathbf{x} = (x_1, \dots, x_n)$ of the tree, which matrices $T = (a_{ij})$ make $p_{x_1 \dots x_n}$ maximal among the coordinates of the distribution p ?

Solution:

Transform the problem to logarithmic coordinates $\log(a_{ij})$, where it becomes linear: $-\log(p_{\mathbf{x}}) \leq -\log(p_\sigma)$ for all σ .

This reduces the problem to that of calculating the convex hull of $-\log(p_\sigma)$ for all labelings σ . The polytopes in the pictures are the result of this calculation.

After computing the polytope, we can immediately

1. Identify **all** labelings which are most likely for some choice of parameters (the vertices of the polytope).
2. Tell which labelings are most likely for given parameters (using linear programming).

Although there are 2^n possible observations, this calculation can be done in polynomial time.

THEOREM: (PACHTER-STURMFELS, [6]) *This convex hull can be calculated in polynomial time in the number of nodes of the model (for a fixed number of parameters) using a geometric version of the sum-product algorithm.*

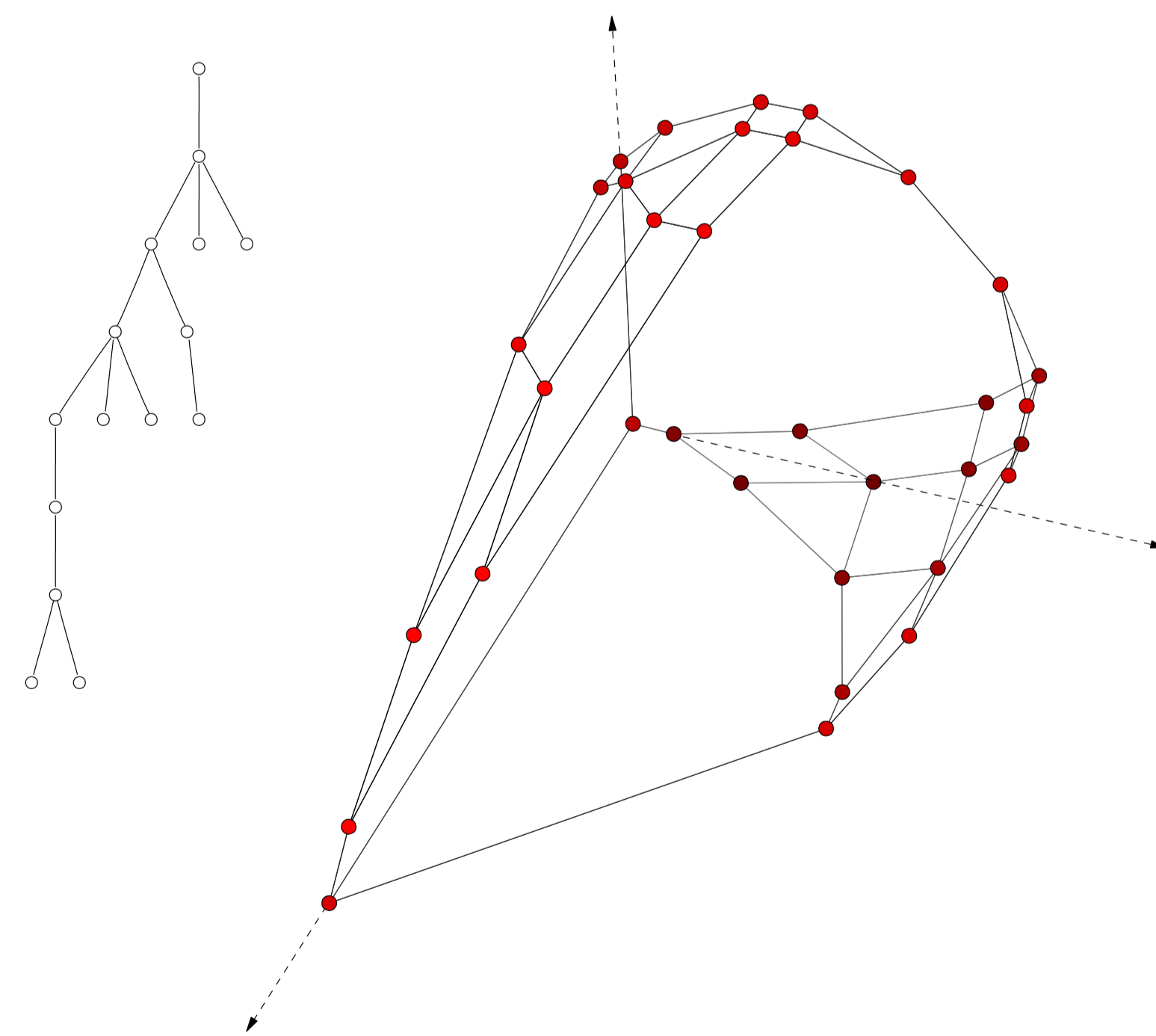
Conjecture:

The number of vertices of the polytope (observations maximal for some choice of the parameters) of any binary tree is bounded by a universal constant.

This conjecture has been verified on all binary trees with at most 15 nodes and selected larger trees, with a maximum of 20 vertices found.

Non binary trees:

However, non-binary trees are not as nice. For example, a tree with 15 nodes can have a polytope with up to 35 vertices. The number of vertices and structure of the polytope depends heavily on the model.

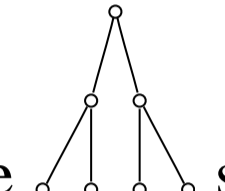


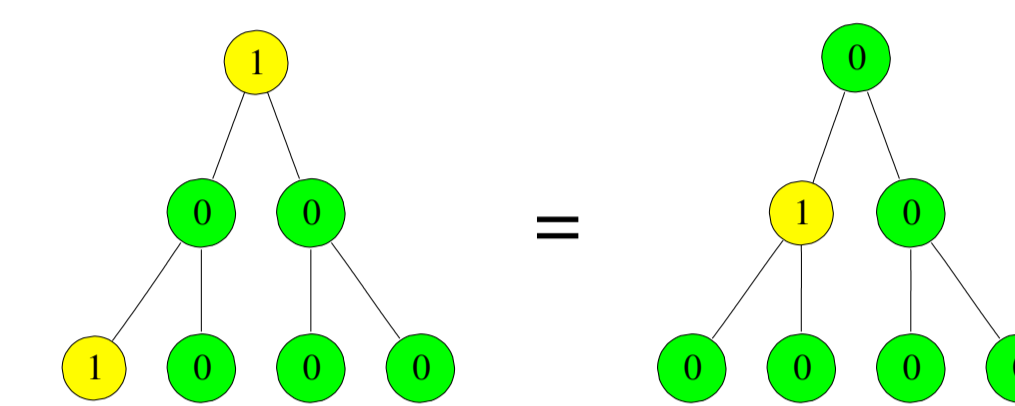
Algebraic Invariants

Definition:

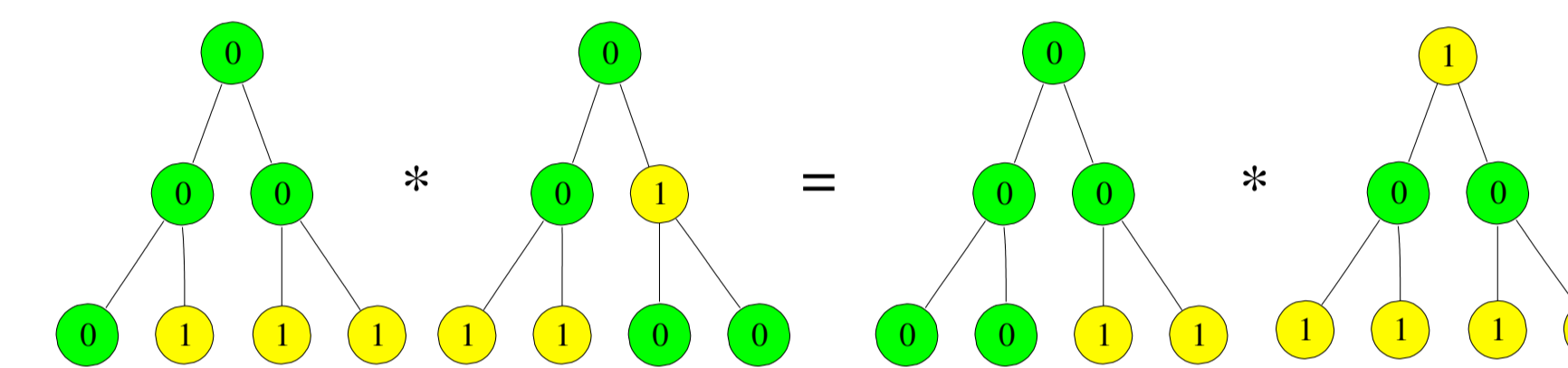
The **algebraic invariants** of a tree T are the parameter independent polynomial relations between the joint probabilities $p_\sigma = P(\mathbf{Y} = \sigma)$.

Example:

The tree  satisfies 441 minimal algebraic invariants. Ninety-six of these are linear invariants such as



The other 345 are quadratic invariants such as



We believe that binary trees require only linear and quadratic invariants.

Computation:

We can compute algebraic invariants very efficiently using **Gröbner bases** with the program 4ti2 [4]. We were able to compute invariants for trees with up to 11 nodes. This is a calculation in 2^{11} variables, which is among the largest ever.

Related Work:

Recently, the algebraic invariants have been described completely for many families of phylogenetic models: the general Markov model (different transition matrices on each edge, all non-leaf nodes hidden) [1], Jukes-Cantor and Kimura models [7], etc.

Conjecture:

The ideal of invariants of a binary tree is generated by polynomials of degree at most 2.

References

- [1] Allman, E. and Rhodes, J. 2003. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 1–33.
- [2] Eriksson, N. 2004. Toric Ideals of Homogeneous Phylogenetic Models, *IS-SAC 2004*, to appear.
- [3] Gusfield, D., Balasubramanian, K., and Naor, D. 1994. Parametric optimization of sequence alignment, *Algorithmica* 12, 312–326.
- [4] R. Hemmecke and R. Hemmecke. 4ti2 version 1.1—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Available at www.4ti2.de, Sept. 2003.
- [5] E. Kuo. 2004. Viterbi sequences and polytopes, eprint: [arXiv:math.CO/0401342](https://arxiv.org/abs/math.CO/0401342)
- [6] Pachter, L. and Sturmfels, B. 2003. Tropical Geometry of Statistical Models, eprint: [arXiv:q-bio.QM/0311009](https://arxiv.org/abs/q-bio.QM/0311009)
- [7] Sturmfels, B. and Sullivan, S. 2004. Toric ideals of phylogenetic invariants, eprint: [arXiv:q-bio.PE/0402015](https://arxiv.org/abs/q-bio.PE/0402015)
- [8] Waterman, M., Eggert, M. and Lander, E. 1992. Parametric sequence comparisons, *Proc. Natl. Acad. Sci. USA* 89:6090–6093.