# Homogeneous Phylogenetic Models: Invariants and Parametric Inference

**Nicholas Eriksson** [1]

**Keywords:** Phylogenetic trees, parametric inference, phylogenetic invariants.

We study the model on phylogenetic trees in which each node is a binary, observable random variable $Y_i$ and the transition probabilities are given by the same matrix $A = \left(\begin{smallmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{smallmatrix}\right)$ on each edge. We write the joint probabilities as $p_{\sigma_1\sigma_2\ldots\sigma_n} := P(\mathbf{Y} = \sigma)$. Let $\rho(i)$ denote the parent of node $i$ in the tree.

The homogeneous Markov model is a fundamental model for statistics. We believe that an understanding of the homogeneous model will lead to more general results. For example, by adding nodes to subdivide edges, we can think of the homogeneous model as a discrete approximation to a continuous Markov model. We are interested in two questions from [4] that are of fundamental importance to the use of statistical models in biology.

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

Problem 1 has been studied in [3, 6] and shown to be of importance for sequence alignment problems. To solve Problem 1, first we write the joint probabilities in terms of the parameters $a_{00}, a_{01}, a_{10}, a_{11}$

$$p_{\sigma_1\sigma_2\ldots\sigma_n} = a_{\sigma_{\rho(2)}\sigma_2} a_{\sigma_{\rho(3)}\sigma_3} \cdots a_{\sigma_{\rho(n)}\sigma_n}.$$

That is, the probability of observing $\sigma$ is the product of the $a_{ij}$ that correspond to the transitions on the edges of the tree. Next, transform to logarithmic coordinates $b_{ij} = -\log(a_{ij})$. The condition that $p_{\sigma_1\ldots\sigma_n}$ is maximal among the coordinates of $p$ becomes the linear system of inequalities

$$b_{\sigma_{\rho(2)}\sigma_2} + \cdots + b_{\sigma_{\rho(n)}\sigma_n} \geq b_{l_{\rho(2)}l_2} + \cdots + b_{l_{\rho(n)}l_n} \qquad \text{for all } (l_1, \ldots l_n) \in \{0,1\}^n.$$

The set of solutions to these inequalities forms a polyhedral cone. If the cone is full dimensional, then $\sigma_1, \ldots, \sigma_n$ is the most likely observation for some choice of the parameters. Such a sequence is called a *Viterbi sequence*. The collection of the cones of all Viterbi sequences is the normal fan of the *Viterbi polytope*, $P_T$. This polytope is three-dimensional and has one vertex for every Viterbi sequence. Given this polytope, we can quickly solve Problem 1 for any $\sigma_1, \ldots, \sigma_n$. Our main result is an explicit description of the polytope for a class of binary trees. For an example, see Figure 1.

**Theorem 1** *If $T$ is a binary tree with $n > 3$ nodes in which all leaves are an odd distance from the root, then there are exactly 8 Viterbi sequences. Furthermore, the polytope $P_T$ has the same combinatorial structure for all such trees.*

Problem 2 is solved by computing the ideal of *polynomial invariants* among the probabilities $p_{\sigma_1\ldots\sigma_n}$. The invariants vanish for a given distribution $(p_{i_1\ldots i_n})$ essentially when that distribution comes from our model. Therefore, invariants have been used in phylogenetics to identify good trees for aligned sequences, see [1, 2]. In the observed, homogeneous Markov

[1]Department of Mathematics, University of California, Berkeley, CA 94720-3840 E-mail: eriksson@math.berkeley.edu

model, the invariants can be computed using the theory of Gröbner bases of toric ideals (see [5]). We are able to calculate the ideal of invariants for trees with 11 nodes. These are computations in 2048 indeterminants, which we believe to be the largest number of indeterminants ever in a Gröbner basis calculation. We conjecture that binary trees require only linear and quadratic generators for the ideal of invariants.

**Example 1** Let $T$ be the path with 4 nodes. Then the ideal of invariants has 32 minimal generators. Four generators are linear (e.g., $p_{0100} - p_{0010}$), twenty-four are quadratic (e.g., $p_{0001} \cdot p_{0010} - p_{0000} \cdot p_{0101}$), and four are cubic (e.g., $p_{1000} \cdot p_{1100} \cdot p_{1111} - p_{0000} \cdot p_{1110}^2$). We believe that the ideal of invariants of a path is always generated by these three types of relations.

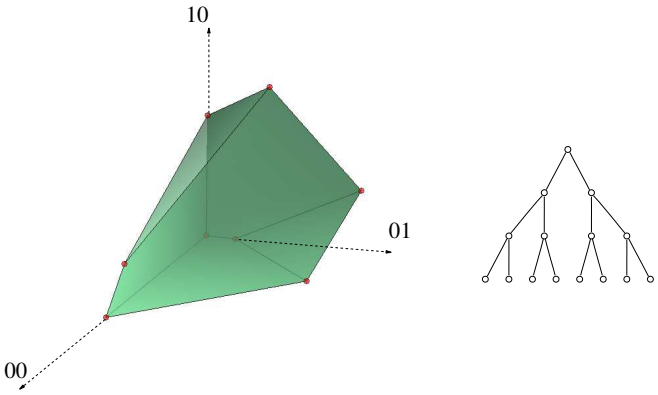


Figure 1: Let $T$ be the pictured tree with 15 nodes. By Theorem 1, since all leaves are at an odd distance from the root, exactly 8 of the $2^{15} = 32768$ possible observations are Viterbi sequences and therefore the polytope $P_T$ has 8 vertices. The polytope is displayed with the $x$-coordinate counting $0 \to 0$ transitions, the $y$-coordinate counting $0 \to 1$ transitions and the $z$-coordinate counting $1 \to 0$ transitions. For example, the front left vertex $(14, 0, 0)$ corresponds to the all zero observation, which is Viterbi sequence with the parameters $a_{00} = 1$, $a_{01} = 0$, $a_{10} = 1$, $a_{11} = 0$.

# References

[1] Allman, E. and Rhodes, J. 2003. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 1–33.

[2] Cavender, J. and Felsenstein, J. 1987. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71.

[3] Gusfield, D., Balasubramanian, K., and Naor, D. 1994. Parametric optimization of sequence alignment, *Algorithmica* 12, 312–326.

[4] Pachter, L. and Sturmfels, B. 2003. The geometry of statistical models for biological sequences, eprint: `arXiv q-bio.QM/0311009`

[5] Sturmfels, B. 1996. *Gröbner bases and convex polytopes*, volume 8 of *University Lecture Series*. American Mathematical Society, Providence, RI.

[6] Waterman, M., Eggert, M. and Lander, E. 1992. Parametric sequence comparisons, *Proc. Natl. Acad. Sci. USA* 89:6090–6093.