

Algebraic Invariants of Mutagenetic Trees

Niko Beerenwinkel, Mathias Drton, Lior Pachter, Bernd Sturmfels¹

Keywords: mutagenetic tree, oncogenetic tree, algebraic invariants, graphical models

1 Introduction.

Mutagenetic trees (also known as *oncogenetic trees*) are a class of graphical models for evolutionary processes that are characterized by the accumulation of genetic changes [3]. They have been used to model the progression of several types of cancer as ordered accumulations of chromosomal alterations. Likewise, the evolution of human immunodeficiency virus under antiviral drug therapy has been described as the accumulation of resistance-conferring mutations [1].

Here we investigate the algebraic geometry of the statistical model induced by a mutagenetic tree [2]. We show that the algebraic invariants of the model have a nice combinatorial structure and can be generated efficiently.

2 Mutagenetic trees.

A mutagenetic tree T for n genetic events is a connected branching on $\{0, \dots, n\}$ rooted at 0 (Fig. 1(a)). Each vertex $v \neq 0$ represents the binary random variable X_v that indicates the occurrence of event v . We associate probability parameters θ_v with the edges of T to obtain the directed acyclic graphical model \mathcal{T} with conditional probability matrices

$$(P(X_v = a \mid X_{\text{pa}(v)} = b))_{a,b=0,1} = \begin{pmatrix} 1 & 0 \\ 1 - \theta_v & \theta_v \end{pmatrix},$$

where $\text{pa}(v)$ denotes the parents of v in T . The first row of this matrix imposes the constraint that an event can occur only if all of its ancestor events have already occurred. We call an observation x of $X = (X_1, \dots, X_n)$ *compatible* with T if there are parameters θ such that T generates x with positive probability. The states compatible with a mutagenetic tree T form a finite distributive lattice $(\mathcal{C}(T), \vee, \wedge)$, where \vee and \wedge denote the componentwise maximum and minimum operator, respectively (Fig. 1(b)).

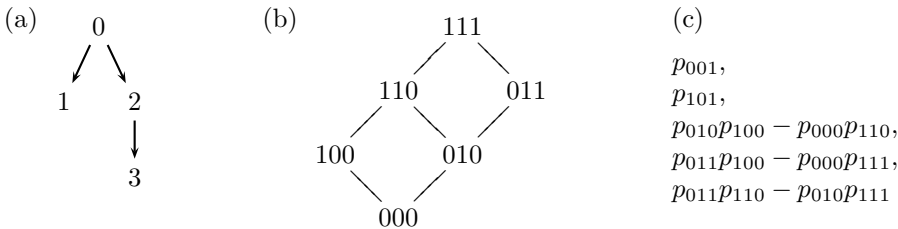


Figure 1: Algebraic invariants of a mutagenetic tree model: (a) Mutagenetic tree T on $n = 3$ events; (b) Induced lattice of compatible states; (c) Gröbner basis for the ideal of invariants.

¹Department of Mathematics, University of California, Berkeley, CA 94720-3840.
E-mail: {niko, drton, lpachter, bernd}@math.berkeley.edu

The set $\mathcal{C}(T)$ can be generated recursively as follows. Let $T(v)$ be the subtree of T rooted at v and denote by $\text{ch}(v)$ the set of children of v in T . If we set

$$\mathcal{C}_v = \begin{cases} \{0, 1\} & \text{if } v \text{ is a leaf,} \\ \{0_{V(T(v))}\} \cup \left(\{1_v\} \times \prod_{u \in \text{ch}(v)} \mathcal{C}_u \right) & \text{else,} \end{cases}$$

then the Cartesian product $\prod_{u \in \text{ch}(0)} \mathcal{C}_u$ equals $\mathcal{C}(T)$.

3 Algebraic Invariants.

The statistical model \mathcal{T} is an n -dimensional algebraic variety in the $2^n - 1$ dimensional probability simplex given by the image of the polynomial map that maps the parameters θ to the model probabilities $p_i(\theta)$, $i \in \mathcal{I} = \{0, 1\}^n$. The algebraic invariants of \mathcal{T} are the polynomials in the polynomial ring $\mathbb{R}[p_i, i \in \mathcal{I}]$ that vanish on \mathcal{T} . The ideal they form is characterized by the following theorems (Fig. 1(c)).

Theorem 1. *The ideal of invariants of a mutagenetic tree \mathcal{T} is generated by the following polynomials:*

- the monomials p_i , i incompatible with T ,
- the squarefree quadratic binomials $p_i p_j - p_{i \vee j} p_{i \wedge j}$, i and j compatible with T , and
- the sum $\sum_{i \in \mathcal{I}} p_i - 1$.

Theorem 2. *Fix the lexicographic order in \mathcal{I} and the reverse lexicographic order in $\mathbb{R}[p_i, i \in \mathcal{I}]$. Then the following polynomials form a reduced Gröbner basis for the ideal generated by the homogeneous invariants of \mathcal{T} :*

- p_i , i incompatible with T , and
- $\underline{p_i p_j} - p_{i \vee j} p_{i \wedge j}$, i, j compatible with T and $(i \wedge j) < i < j < (i \vee j)$.

The underlined terms are the leading monomials according to the fixed term order.

In summary, the algebraic invariants of mutagenetic trees can be constructed combinatorially without the use of computer algebra techniques such as elimination and Gröbner basis computation. The computational complexity of generating the Gröbner basis in Theorem 2 is only linear in the size of the output. The length of this basis is $O(4^n)$, but the specific combinatorial structure suggests more compressed representations. The complete knowledge of the invariants provides a solid basis for model selection, parameter estimation, and statistical inference.

References

- [1] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. In *Proc. 8th Ann. Int. Conf. on Res. in Comput. Biol. (RECOMB '04)*, 27–31 March 2004, San Diego, CA, pages 36–44, 2004, to appear in *J. Comp. Biol.*
- [2] N. Beerenwinkel and M. Drton. Mutagenetic tree models. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for Computational Biology*, chapter 14. Cambridge University Press, Cambridge, UK, 2005, to appear.
- [3] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C.H. Papadimitriou, and A.A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comp. Biol.*, 6(1):37–51, 1999.