Algebraic Phylogenetics

Sepehr Akbari

4/27/2026

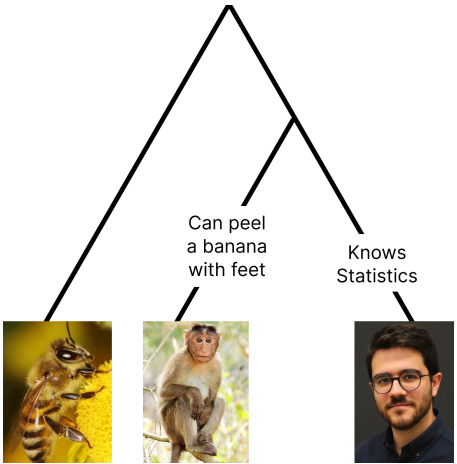
Department of Math & CS, Lake Forest College

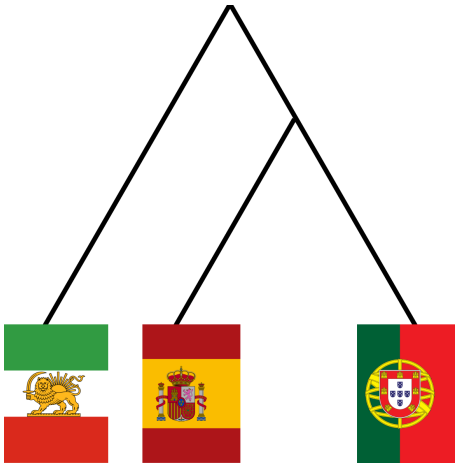
Motivated	Biologically
Viewed	Statistically
Solved	Algebraically

Biological Motivation

Phylogenetic Trees

In biology, we view the evolutionary history of species as a directed tree graph.





Statistical View

Statistical Model

We can view the leaves of the tree as $\vec{S}_1, \vec{S}_2, \ldots, \vec{S}_n$ RVs, and the internal (latent) nodes as $\vec{H}_1, \vec{H}_2, \ldots, \vec{H}_m$ RVs.

Each $\vec{S}_i$ is a vector of DNA bases, where each position (site) is a random variable X that takes one of four states in

$$\Sigma = \{A,\ C,\ T,\ G\}$$

For example, at a single site in the genome, we might have $X_{\text{Bee}} = A$, $X_{\text{Monkey}} = C$, and $X_{\text{Me}} = C$.

Goal. Given only the observed data at the leaves, we want to infer the correct structure (topology) of the tree when adding a new species.

Modeling Mutations

We model these transition probabilities (mutations) as a **Discrete Markov Process**.

For every directed edge $g = (u, v)$ (from parent u to child v), there is a 4×4 matrix $M_{ij}^{(g)}$ that gives the probability that a parent with base i mutates into a child with base j:

$$M_{ij}^{(g)} = \mathbb{P}(X_v = j \mid X_u = i)$$

We desire to compare the likelihood of the observed data under different tree topologies, which requires us to estimate all $M^{(g)}$ matrices.

Can we just use MLEs...?

MLE Formulation

Let our parameters be $\theta = (T, M)$, where T is a specific tree topology and M represents all the edge matrices.

For a single site k, the probability of observing the specific leaf data requires summing (marginalizing) over all possible states of the hidden internal nodes:

$$\mathbb{P}(\text{site } k \mid \theta) = \sum_{H} \pi_\rho \cdot M^{(g_1)} \cdot M^{(g_2)} \dots M^{(g_E)}$$

where π_ρ is the probability distribution at the root.

For a DNA sequence of length N, the full Likelihood function is the product over all independent sites:

$$\mathcal{L}(\theta \mid \text{Data}) = \prod_{k=1}^{N} \left(\sum_{H} \pi_\rho \cdot M^{(g_1)} \dots M^{(g_E)} \right)$$

MLE Failure

There are three major issues with this MLE formulation:

Not Closed-Form. $\ell = \log \mathcal{L}$ cannot be simplified due to the sum. Taking derivatives yields a system with no closed-form solution.

Local Minima. Because of the hidden nodes, the likelihood surface is non-convex. Numerical optimization frequently gets trapped in local maxima.

Combinatorial Explosion. The number of possible tree topologies grows factorially with the number of species. For 50 species, there are more trees than atoms in the universe.

We need a shift in perspective.

Algebraic Solution

Polynomial Representation

Recall our joint probability for a specific observed leaf pattern, $x = (x_1, \ldots, x_n)$. We marginalize over all possible states of the hidden internal nodes $h = (h_1, \ldots, h_m)$:

$$p_x = \sum_{h \in \Sigma^m} \pi_\rho \prod_{(u,v) \in E} M^{(u,v)}_{h_u, x_v}$$

For a simple 3-leaf tree with one hidden root, observing $x = (A, C, C)$ expands to:

$$p_{ACC} = \pi_\rho(A)\left(M^{(g_1)}_{AA} M^{(g_2)}_{AC} M^{(g_3)}_{AC}\right) + \pi_\rho(C)\left(M^{(g_1)}_{CA} M^{(g_2)}_{CC} M^{(g_3)}_{CC}\right) + \ldots$$

Observe that p_x is a *multivariate polynomial* where the variables are the unknown transition probabilities $M^{(g)}_{ij}$.

Question. Can we find relationships between these leaf probabilities, p_x, that do *not* depend on the unknown matrices $M^{(g)}$?

Algebraic Signatures

Fact. For any specific tree topology T, there exists a set of polynomials that will *always* evaluate to exactly zero on the true leaf probabilities P, regardless of the actual transition matrices:

$$f(p_{AA\ldots}, \ldots, p_{GG\ldots}) = 0$$

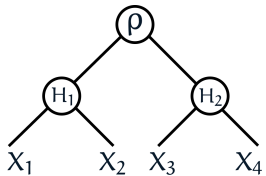
The set of all such polynomials for a tree T is denoted $\mathcal{I}_T$. This set is unique to T and serves as its algebraic "signature."

If we evaluate a polynomial from $\mathcal{I}_T$ using our observed data and it does *not* equal zero, we can reject T as the true topology.

Constructing $\mathcal{I}_T$

How do we find the polynomials in $\mathcal{I}_T$?

Consider a tree with 4 species, and us wanting to find the invariants $p_{x_1x_2x_3x_4}$.



$$
\begin{aligned}
p_{x_1x_2x_3x_4} &= \sum_{\rho\in\Sigma}^{T}\sum_{H_1\in\Sigma}^{T}\sum_{H_2\in\Sigma}^{T} \pi_\rho \cdot M^{(g_1)}_{\rho,H_1} M^{(g_2)}_{\rho,H_2} M^{(g_3)}_{H_1,x_1} M^{(g_4)}_{H_2,x_3} M^{(g_5)}_{H_1,x_2} M^{(g_6)}_{H_2,x_4} \\
&= \sum_{\rho\in\Sigma}^{T} \pi_\rho \cdot \left[\left(\sum_{H_1\in\Sigma}^{T} M^{(g_1)}_{\rho,H_1} M^{(g_3)}_{H_1,x_1} M^{(g_5)}_{H_1,x_2} \right) \cdot \left(\sum_{H_2\in\Sigma}^{T} M^{(g_2)}_{\rho,H_2} M^{(g_4)}_{H_2,x_3} M^{(g_6)}_{H_2,x_4} \right) \right]
\end{aligned}
$$

Flattening

We can "flatten" this 4D tensor into a 2D matrix by partitioning the leaves into two sets, say $L_1 = \{X_1, X_2\}$ and $L_2 = \{X_3, X_4\}$. We can construct a 16×16 flattening matrix, denoted $\mathcal{F}_{L_1|L_2}(P)$ as:

$$\mathcal{F}_{L_1|L_2}(P) = \begin{pmatrix} p_{AAAA} & p_{AAAC} & \cdots & p_{AACA} & p_{AACC} & \cdots & p_{AAGG} \\ p_{ACAA} & p_{ACAC} & \cdots & p_{ACCA} & p_{ACCC} & \cdots & p_{ACGG} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_{GTAA} & p_{GTAC} & \cdots & p_{GTCA} & p_{GTCC} & \cdots & p_{GTGG} \\ p_{GGAA} & p_{GGAC} & \cdots & p_{GGCA} & p_{GGCC} & \cdots & p_{GGGG} \end{pmatrix}$$

But why do we care about this flattening? Its rank!

Observe that for a fixed ρ, we simply have a rank-1 matrix:

$$\mathbb{P}(X_1, \ldots, X_4 \mid \rho) = \left(\sum_{H_1 \in \Sigma}^{T} M^{(g_1)}_{\rho,H_1} M^{(g_3)}_{H_1,x_1} M^{(g_5)}_{H_1,x_2} \right) \cdot \left(\sum_{H_2 \in \Sigma}^{T} M^{(g_2)}_{\rho,H_2} M^{(g_4)}_{H_2,x_3} M^{(g_6)}_{H_2,x_4} \right)$$

where the terms are a 16×1 column vector and a 1×16 row vector, respectively. Hence, because we have a sum of 4 rank-1 matrices (one for each ρ):

$$\text{Rank}(\mathcal{F}_{L_1|L_2}(P)) \leq 4$$

Recall that if you select $(r + 1)$ rows from a matrix of rank r, those rows are linearly dependent. Therefore:

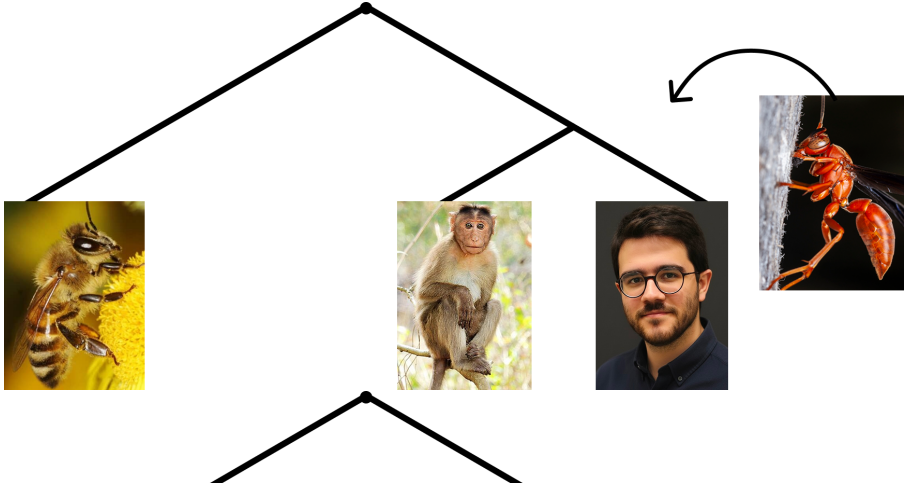
$$\det(\mathcal{G}) = 0, \quad \forall G_{5\times 5\geq} \in \mathcal{F}_{L_1|L_2}(P)$$

These determinants are the elements of $\mathcal{I}_T$. Polynomials in terms of p_x that vanish for the true tree topology, regardless of the transition probabilities.

Why is any of this useful?

Validating Tree Topologies

Consider our original tree. We want to add a new species, say a wasp, and determine where it fits in the tree.



Validating Tree Topologies

To do this mathematically, we follow the algebraic framework.

(i) **Collect the DNA samples of the species.**
say we have 1000 sites in the genome, so we have a dataset of 4×1000 bases.

(ii) **Compute empirical probabilities p_x.**
say you count (A, A, A, A) 30 times, so $p_{AAAA} = 30/1000 = 0.03$.

(iii) **Evaluate the invariants in $\mathcal{I}_T$ on the empirical distribution $\hat{P}$.**
say we flatten $\hat{P}$ into $\mathcal{F}_{\{\text{Bee,Sepehr}\} \mid \{\text{Monkey,Wasp}\}}(\hat{P})$.

(iv) **If the invariants evaluate to zero, then T is a valid topology for the data. If not, we reject T and try a different topology.**
we compute the determinant of all minors, but we have to reject the topology since $\det(\mathcal{G}) \neq 0 + \epsilon$ for some $\mathcal{G}_{5\times5}$. We try another topology.

What Else?

The broader usage of this framework is in machine learning, to evaluate the **identifiability** of a model.

Identifiability asks whether we can uniquely recover the true parameters of a model from the observed data.

$$\theta_1 \neq \theta_2 \implies \mathbb{P}(\vec{x}_i \mid \theta_1) \neq \mathbb{P}(\vec{x}_i \mid \theta_2)$$

For a complex model, especially with millions of parameters, we want to know if the true parameters can be uniquely identified. If a model is unidentifiable, no amount of data will allow us to learn the ground truth.

Algebraic statistics provides tools to study identifiability.

References

[1] Elizabeth S. Allman and John A. Rhodes. *"Phylogenetic ideals and varieties for the general Markov model"*. In: (2006).

[2] Nicholas Eriksson et al. *"Phylogenetic Algebraic Geometry"*. In: (2004).

[3] Sara Jamshidi. *Algebraic Geometry of Tree Tensor Network States*. Comprehensive Talk. 2013.

Thank You!

Questions?

Sepehr Akbari
akbaris79@lakeforest.edu
Department of Math & CS, Lake Forest College

Algebraic Tools

Phylogenetic Invariants

A polynomial that evaluates to zero for a specific tree topology T, no matter the transition matrices, is called a **Phylogenetic Invariant**:

$$f(p_{AA\ldots}, \ldots, p_{GG\ldots}) = 0$$

In Algebraic Geometry, the set of all such invariant polynomials for a tree T forms an **Ideal**, denoted $\mathcal{I}_T$.

$$\mathcal{I}_T \coloneqq \{f \in \mathbb{R}[p_x] \mid f(P) = 0,\ \forall P \in V(\mathcal{I}_T)\}$$

where $V(\mathcal{I}_T)$ is a set of all probability distributions that satisfy the invariants in $\mathcal{I}_T$, called the **Variety** of the ideal.

You should think of the Ideal as the algebraic "signature" of the tree.

Computation

Instead of computing the determinant of every minor in our flattening, we can instead use a **Gröbner basis** of the ideal $\mathcal{I}_T$ to determine if a distribution P is consistent with tree T.

A Gröbner basis is a special generating set of an ideal that allows us to efficiently check if a polynomial belongs to the ideal.

If we can express the polynomial f as a linear combination of the Gröbner basis elements, then f belongs to $\mathcal{I}_T$.

Thank You!

Questions?

Sepehr Akbari
akbaris79@lakeforest.edu
Department of Math & CS, Lake Forest College