

Phylogenetic Models and Algebra

Isaac Newton Institute

Spitalfields Day

December 6, 2007

John A. Rhodes

Department of Mathematics and Statistics

University of Alaska Fairbanks 

Mathematical aspects of Phylogenetics:

- Modeling
- Combinatorics
- Algebra
- ??
- Statistics
- Algorithms
- Geometry

I: Simple model of biological sequence evolution

Suppose sequences are composed of two letters R,Y.

A : RRYRYRYYYRYRYR...



D : RRRRYYYRYYYRYYYR...

Ancestral sequence: Generated by (biased) coin tosses

$$\text{Prob}(R) = \pi_R \quad \text{Prob}(Y) = \pi_Y$$

Ex: $\boldsymbol{\pi} = (\pi_R \ \pi_Y) = (.6 \ .4) \rightsquigarrow$

A : RRYRYRYYYRRYRRYRYRRRY...

Evolution of sequence:

A : RRYRYRYYYRYRYR...



D : RRRRYYYRYYYRYYYR...

Markov matrix gives probabilities of changes at each site:

$$M = \begin{pmatrix} m_{RR} & m_{RY} \\ m_{YR} & m_{YY} \end{pmatrix}$$

where $m_{ij} = \text{Prob}(i \rightarrow j)$

What does the model predict for probabilities of *patterns* in aligned sequences?

A : RRYRYRYYYRYRYR...

↓

D : RRRRYYRYYYRYYYR...

$$p_{ij} = \text{Prob}(A = i, D = j) = \pi_i m_{ij}$$

$$\begin{pmatrix} p_{RR} & p_{RY} \\ p_{YR} & p_{YY} \end{pmatrix} = \begin{pmatrix} \pi_R & 0 \\ 0 & \pi_Y \end{pmatrix} \begin{pmatrix} m_{RR} & m_{RY} \\ m_{YR} & m_{YY} \end{pmatrix}$$

$$P = \text{diag}(\boldsymbol{\pi})M.$$

From data, we can determine (estimate) P :

A : RRRRYRYRY...

D : YRRYRYRYRR...

$$P = \begin{pmatrix} p_{RR} & p_{RY} \\ p_{YR} & p_{YY} \end{pmatrix} \approx \begin{pmatrix} .4 & .2 \\ .2 & .2 \end{pmatrix}.$$

Does P uniquely determine the parameters π , M ?

(Does data retain all information on the process?) YES

$$\pi = (.6 \ .4), \quad M = \begin{pmatrix} .667 & .333 \\ .5 & .5 \end{pmatrix}$$

Continuous time process:

$$Q = \begin{pmatrix} -r & r \\ s & -s \end{pmatrix}$$

$r =$ rate at which $R \rightarrow Y$, $s =$ rate at which $Y \rightarrow R$

then

$$M = \exp(Qt), \quad t = \text{elapsed time}$$

- Continuous-time models are embedded in algebraic models.
- Algebraic models can be understood through additional tools.

But this is too simple —

- 4 bases A,C,T,G \rightsquigarrow 4×4 matrices, etc.
- evolution proceeds down a tree, with branches
- we don't know ancestral sequence
- we have many sequences, from leaves of a tree
- we don't know the tree
- we don't know any of the numerical parameters
- we need more elaborate models for biological realism

Statistical framework for using a model:

Data \approx prediction of some model using some parameters

↑

?

How do we choose parameters for which this holds?

E.g.,

Maximum Likelihood: choose parameters that maximize

Prob(data | parameters)

II: Identifiability of models

Fundamental question:

Given a model of sequence evolution, does the resulting distribution of observable data uniquely determine all model parameters?

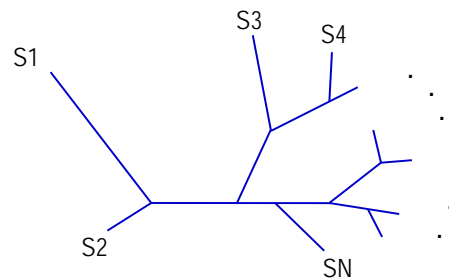
Is the model *identifiable*?

More informally,

From 'perfect' data, is it possible to infer the details of the process leading to it?

General model along a tree T :

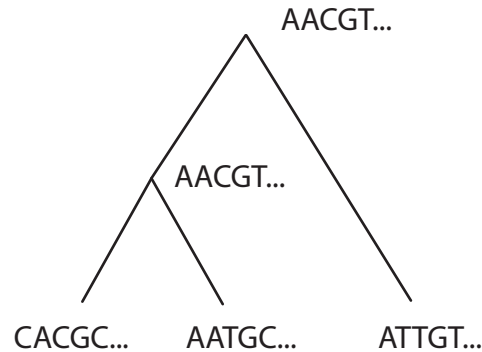
Fix an n -taxon tree T , κ states at each node,



$\kappa = 4$ (DNA), $\kappa = 2$ (R/Y), or $\kappa = 20$ (proteins)

$$\text{parameters} = \begin{cases} \text{tree} \\ \text{root distribution vector } \pi \\ \text{Markov matrix on each edge } M_e \end{cases}$$

More specifically,

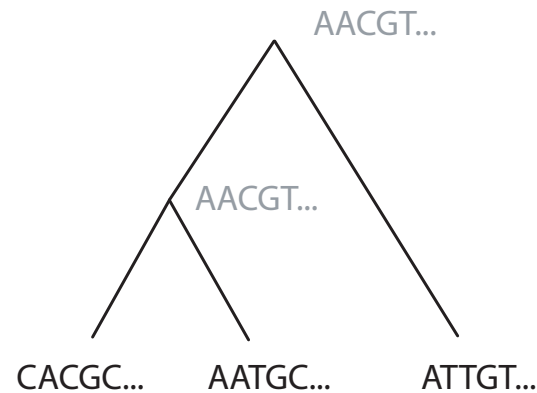


Description of process at **a single site**:

- Bases $1, 2, \dots, \kappa$ (For DNA, $A, C, G, T \rightsquigarrow 1, 2, 3, 4$)
- Bases at root occur with probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_\kappa)$; $\sum \pi_i = 1$.
- On each edge e , Markov matrix M_e give probs. of base substitutions,

$$M_e(i, j) = \text{Prob}(i \rightarrow j)$$

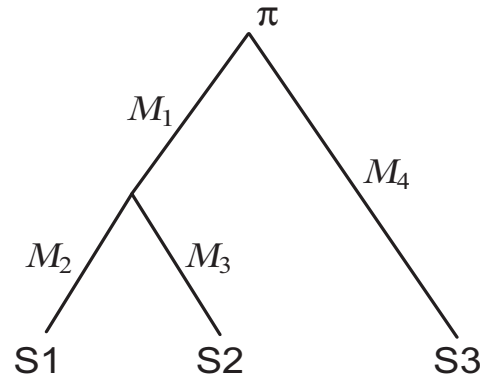
This is the **general Markov model** (GM) on the tree T .



Note:

- Multiple sites, i.i.d.
- We observe states only of living taxa (leaves); states at all internal nodes are *hidden*.
- Given an ancestral state at any node, processes on descending edges are independent.

Given $T, \pi, \{M_e\}$,

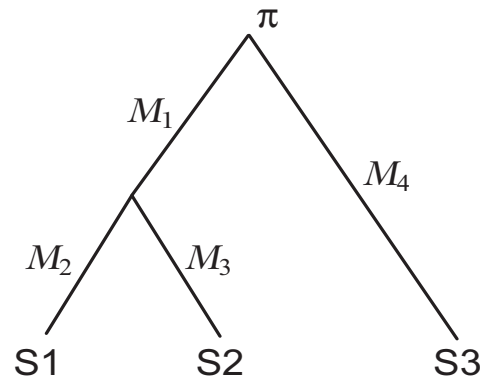


compute distribution P of patterns of bases at leaves:

E.g., $GAA \rightsquigarrow 311$,

$$p_{311} = \sum_{i=1}^4 \sum_{j=1}^4 \pi_i M_1(i, j) M_2(j, 3) M_3(j, 1) M_4(i, 1)$$

$P = (p_{lmn})$, a 3-dim tensor / array / table



$$P(l, m, n) = p_{lmn} = \sum_{i=1}^4 \sum_{j=1}^4 \pi_i M_1(i, j) M_2(j, l) M_3(j, m) M_4(i, n)$$

- P is $4 \times 4 \times 4$,
- $P(l, m, n)$ are polynomial (16 quintic terms)
- Form of polynomial encodes the tree topology,
- Larger trees produce higher-degree terms, and more terms,

Q: Given P , a pattern distribution predicted by a model, can we identify

1) the tree?

2) the numerical parameters?

If answer is NO, then statistical inference is problematic.

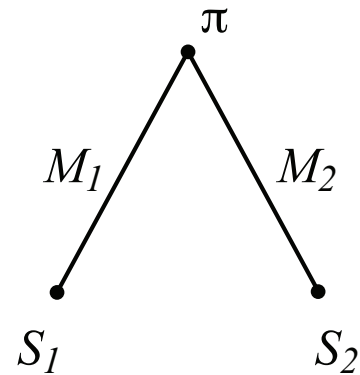
Note: Polynomial maps are generally many-to-one.

There are known cases of models similar to phylogenetic ones (latent class models) that are *not* identifiable.

A: For the model discussed so far, essentially yes.

For more complicated models, many questions are open.

Ex:



Aligned sequences described by

$$\text{Prob}(S_1 = i, S_2 = j) = P_{ij} = \sum_{l \in \{R, Y\}} \pi_l M_1(l, i) M_2(l, j)$$

or
$$P = M_1^T \text{diag}(\boldsymbol{\pi}) M_2.$$

Does P determine π, M_1, M_2 ? NO

e.g.,

$$\begin{aligned} P &= \begin{pmatrix} .9 & .1 \\ .1 & .9 \end{pmatrix}^T \begin{pmatrix} .5 & 0 \\ 0 & .5 \end{pmatrix} \begin{pmatrix} .9 & .1 \\ .1 & .9 \end{pmatrix} \\ &= \begin{pmatrix} .41 & .09 \\ .09 & .41 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^T \begin{pmatrix} .5 & 0 \\ 0 & .5 \end{pmatrix} \begin{pmatrix} .82 & .18 \\ .18 & .82 \end{pmatrix} \end{aligned}$$

Non-uniqueness of this matrix factorization

↪ ancestor cannot be recovered

by statistics alone.

Larger trees require we go beyond matrix algebra

A broadening of perspective ...

When specifying a model of molecular evolution on a fixed tree T , the joint distribution map defines a **polynomial** parameterization:

Parameter Space \longrightarrow Joint Distribution space

$$(\boldsymbol{\pi}, M_e) \longmapsto P$$

The realm of mathematics that studies polynomial maps is **algebraic geometry**.

Stochastic setting:

$$\begin{aligned}\phi_T &: [0, 1]^N \longrightarrow [0, 1]^{\kappa^n} \\ (\boldsymbol{\pi}, \{M_e\}) &\longmapsto P\end{aligned}$$

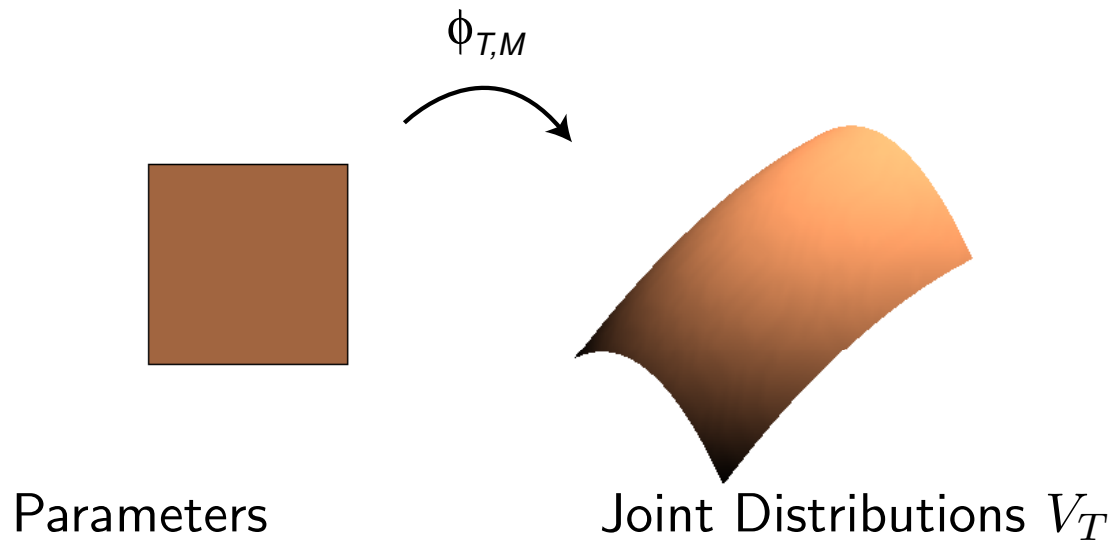
Geometric Setting:

$$\begin{aligned}\phi_T &: \mathbb{C}^N \longrightarrow \mathbb{C}^{\kappa^n} \\ (\boldsymbol{\pi}, \{M_e\}) &\longmapsto P\end{aligned}$$

Each tree T is associated to a parameterized surface $V_T = \overline{\text{Im}(\phi_T)}$, the *phylogenetic variety* V_T .

(The biologically relevant joint distributions are a subset of the points on V_T .)

Geometric view:



For each tree $T \in \mathcal{T}$,

the map $\phi_{T,\mathcal{M}}$ defines a 'surface' $V_T = \overline{\text{Im}(\phi_{T,\mathcal{M}})}$.

'surface' = algebraic variety

The tree parameter for \mathcal{M} is identifiable if $V_{T_i} \cap V_{T_j} = \emptyset$ for all $i \neq j$.



Tree identifiability fails if two such varieties intersect at any points.

However, ...

Limitations on Identifiability of Tree topology:

In fact $V_{T_1} \cap V_{T_2} \neq \emptyset$ (due to edges of length 0, ∞)



Identifiability still holds
for 'most' parameters.

If the intersection is of lower dimension, then the tree is **identifiable for generic parameters**.

Given parameterizations of V_{T_1}, V_{T_2} , how do we show intersection is of lower dimension?

The parameterized phylogenetic variety V_T also has an *implicit description*, as the zero set of polynomials in the *phylogenetic ideal* I_T .

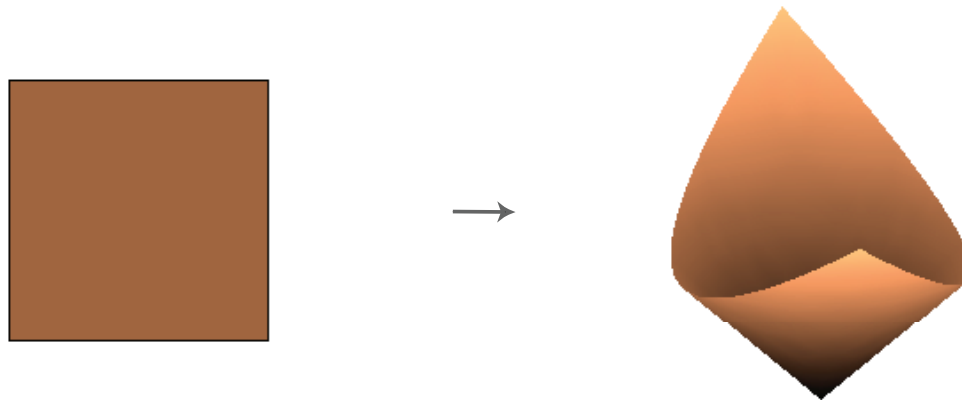
$$P \in V_T \Leftrightarrow f(P) = 0 \text{ for all } f \in I_T$$

Polynomials in I_T are known as *phylogenetic invariants*.

Non-phylogenetic example:

Parameterization:

$$\begin{aligned}\phi : \mathbb{C}^2 &\rightarrow \mathbb{C}^3, \\ (s, t) &\mapsto (t^2 - s^2, 2st, t^2 + s^2),\end{aligned}$$



If $f(x, y, z) = x^2 + y^2 - z^2$, then $f(\phi(s, t)) = 0$.

Implicit description:

$$x^2 + y^2 - z^2 = 0$$

Parameterized vs. implicit descriptions:

Parameterizations are better for:

- producing points on a variety

Implicit descriptions are better for:

- determining if a given point is on a variety
- determining if two varieties are the same
- determining the intersection of two varieties
- determining points of self-intersection

Only one invariant is easy to see – stochastic invariant

$$1 - \sum_{\text{pattern}} p_{\text{pattern}}$$

Typically they are of higher degree and reflect the topology of the tree T and choice of mutation model.

Eg. GM model, $\kappa = 4$, for 3-leaf tree, lowest degree invariants are of degree 5, for example...

$$\begin{aligned}
 f = & -p_{121}p_{133}p_{002}p_{212}p_{322} + p_{121}p_{133}p_{002}p_{222}p_{312} + p_{121}p_{133}p_{202}p_{012}p_{322} \\
 & - p_{121}p_{133}p_{202}p_{022}p_{312} - p_{121}p_{133}p_{302}p_{012}p_{222} + p_{121}p_{133}p_{302}p_{022}p_{212} \\
 & + p_{321}p_{103}p_{012}p_{122}p_{232} - p_{321}p_{103}p_{012}p_{132}p_{222} - p_{321}p_{103}p_{112}p_{022}p_{232} \\
 & + p_{321}p_{103}p_{112}p_{032}p_{222} + p_{321}p_{103}p_{212}p_{022}p_{132} - p_{321}p_{103}p_{212}p_{032}p_{122} \\
 & - p_{321}p_{113}p_{002}p_{122}p_{232} + p_{321}p_{113}p_{002}p_{132}p_{222} + p_{321}p_{113}p_{102}p_{022}p_{232} \\
 & - p_{321}p_{113}p_{102}p_{032}p_{222} - p_{321}p_{113}p_{202}p_{022}p_{132} + p_{321}p_{113}p_{202}p_{032}p_{122} \\
 & + p_{321}p_{123}p_{002}p_{112}p_{232} - p_{321}p_{123}p_{002}p_{132}p_{212} - p_{321}p_{123}p_{102}p_{012}p_{232} \\
 & + p_{321}p_{123}p_{102}p_{032}p_{212} + p_{321}p_{123}p_{202}p_{012}p_{132} - p_{321}p_{123}p_{202}p_{032}p_{112} \\
 & - p_{321}p_{133}p_{002}p_{112}p_{222} + p_{321}p_{133}p_{002}p_{122}p_{212} + p_{321}p_{133}p_{102}p_{012}p_{222} \\
 & - p_{321}p_{133}p_{102}p_{022}p_{212} - p_{321}p_{133}p_{202}p_{012}p_{122} + p_{321}p_{133}p_{202}p_{022}p_{112} \\
 & - p_{323}p_{101}p_{212}p_{022}p_{132} + p_{323}p_{101}p_{212}p_{032}p_{122} + p_{323}p_{111}p_{002}p_{122}p_{232} \\
 & - p_{323}p_{111}p_{002}p_{132}p_{222} - p_{323}p_{111}p_{102}p_{022}p_{232} + p_{323}p_{111}p_{102}p_{032}p_{222} \\
 & + p_{323}p_{111}p_{202}p_{022}p_{132} - p_{323}p_{111}p_{202}p_{032}p_{122} - p_{323}p_{121}p_{002}p_{112}p_{232}
 \end{aligned}$$

$+p_{323}p_{121}p_{002}p_{132}p_{212} + p_{323}p_{121}p_{102}p_{012}p_{232} - p_{323}p_{121}p_{102}p_{032}p_{212}$
 $-p_{323}p_{121}p_{202}p_{012}p_{132} + p_{323}p_{121}p_{202}p_{032}p_{112} + p_{323}p_{131}p_{002}p_{112}p_{222}$
 $-p_{323}p_{131}p_{002}p_{122}p_{212} - p_{323}p_{131}p_{102}p_{012}p_{222} + p_{323}p_{131}p_{102}p_{022}p_{212}$
 $+p_{323}p_{131}p_{202}p_{012}p_{122} - p_{323}p_{131}p_{202}p_{022}p_{112} - p_{223}p_{111}p_{302}p_{022}p_{132}$
 $+p_{223}p_{111}p_{302}p_{032}p_{122} - p_{121}p_{103}p_{012}p_{232}p_{322} - p_{221}p_{103}p_{012}p_{122}p_{332}$
 $+p_{221}p_{103}p_{012}p_{132}p_{322} + p_{221}p_{103}p_{112}p_{022}p_{332} - p_{221}p_{103}p_{112}p_{032}p_{322}$
 $-p_{221}p_{103}p_{312}p_{022}p_{132} + p_{221}p_{103}p_{312}p_{032}p_{122} + p_{221}p_{113}p_{002}p_{122}p_{332}$
 $-p_{221}p_{113}p_{002}p_{132}p_{322} - p_{221}p_{113}p_{102}p_{022}p_{332} + p_{221}p_{113}p_{102}p_{032}p_{322}$
 $+p_{221}p_{113}p_{302}p_{022}p_{132} - p_{221}p_{113}p_{302}p_{032}p_{122} - p_{221}p_{123}p_{002}p_{112}p_{332}$
 $+p_{221}p_{123}p_{002}p_{132}p_{312} + p_{221}p_{123}p_{102}p_{012}p_{332} - p_{221}p_{123}p_{102}p_{032}p_{312}$
 $-p_{221}p_{123}p_{302}p_{012}p_{132} + p_{221}p_{123}p_{302}p_{032}p_{112} + p_{221}p_{133}p_{002}p_{112}p_{322}$
 $-p_{221}p_{133}p_{002}p_{122}p_{312} - p_{221}p_{133}p_{102}p_{012}p_{322} + p_{221}p_{133}p_{102}p_{022}p_{312}$
 $+p_{221}p_{133}p_{302}p_{012}p_{122} - p_{221}p_{133}p_{302}p_{022}p_{112} - p_{223}p_{101}p_{012}p_{132}p_{322}$
 $-p_{223}p_{101}p_{112}p_{022}p_{332} + p_{121}p_{103}p_{212}p_{032}p_{322} + p_{121}p_{103}p_{312}p_{022}p_{232}$
 $-p_{123}p_{101}p_{012}p_{222}p_{332} + p_{123}p_{101}p_{012}p_{232}p_{322} + p_{123}p_{101}p_{212}p_{022}p_{332}$
 $-p_{123}p_{101}p_{212}p_{032}p_{322} - p_{123}p_{101}p_{312}p_{022}p_{232} + p_{123}p_{101}p_{312}p_{032}p_{222}$
 $+p_{123}p_{111}p_{002}p_{222}p_{332} - p_{123}p_{111}p_{002}p_{232}p_{322} - p_{123}p_{111}p_{202}p_{022}p_{332}$
 $+p_{123}p_{111}p_{202}p_{032}p_{322} + p_{123}p_{111}p_{302}p_{022}p_{232} - p_{123}p_{111}p_{302}p_{032}p_{222}$
 $+p_{123}p_{131}p_{002}p_{212}p_{322} - p_{123}p_{131}p_{002}p_{222}p_{312} - p_{123}p_{131}p_{202}p_{012}p_{322}$
 $+p_{123}p_{131}p_{202}p_{022}p_{312} + p_{123}p_{131}p_{302}p_{012}p_{222} - p_{123}p_{131}p_{302}p_{022}p_{212}$
 $-p_{021}p_{103}p_{112}p_{222}p_{332} + p_{021}p_{103}p_{112}p_{232}p_{322} + p_{021}p_{103}p_{212}p_{122}p_{332}$
 $-p_{021}p_{103}p_{212}p_{132}p_{322} - p_{021}p_{103}p_{312}p_{122}p_{232} + p_{021}p_{103}p_{312}p_{132}p_{222}$
 $+p_{021}p_{113}p_{102}p_{222}p_{332} - p_{021}p_{113}p_{102}p_{232}p_{322} - p_{021}p_{113}p_{202}p_{122}p_{332}$
 $+p_{021}p_{113}p_{202}p_{132}p_{322} + p_{021}p_{113}p_{302}p_{122}p_{232} - p_{021}p_{113}p_{302}p_{132}p_{222}$
 $-p_{021}p_{123}p_{102}p_{212}p_{332} + p_{021}p_{123}p_{102}p_{232}p_{312} + p_{021}p_{123}p_{202}p_{112}p_{332}$
 $-p_{021}p_{123}p_{202}p_{132}p_{312} + p_{023}p_{121}p_{202}p_{132}p_{312} + p_{023}p_{121}p_{302}p_{112}p_{232}$
 $+p_{223}p_{101}p_{012}p_{122}p_{332} + p_{223}p_{101}p_{112}p_{032}p_{322} + p_{223}p_{101}p_{312}p_{022}p_{132}$
 $-p_{223}p_{101}p_{312}p_{032}p_{122} - p_{223}p_{111}p_{002}p_{122}p_{332} + p_{223}p_{111}p_{002}p_{132}p_{322}$

$$\begin{aligned}
& +p_{223}p_{111}p_{102}p_{022}p_{332} - p_{223}p_{111}p_{102}p_{032}p_{322} + p_{023}p_{101}p_{112}p_{222}p_{332} \\
& - p_{023}p_{101}p_{112}p_{232}p_{322} - p_{023}p_{101}p_{212}p_{122}p_{332} + p_{023}p_{101}p_{212}p_{132}p_{322} \\
& + p_{023}p_{101}p_{312}p_{122}p_{232} - p_{023}p_{101}p_{312}p_{132}p_{222} - p_{023}p_{111}p_{102}p_{222}p_{332} \\
& + p_{023}p_{111}p_{102}p_{232}p_{322} + p_{023}p_{111}p_{202}p_{122}p_{332} - p_{023}p_{111}p_{202}p_{132}p_{322} \\
& - p_{023}p_{111}p_{302}p_{122}p_{232} + p_{023}p_{111}p_{302}p_{132}p_{222} + p_{023}p_{121}p_{102}p_{212}p_{332} \\
& - p_{023}p_{121}p_{102}p_{232}p_{312} - p_{023}p_{121}p_{202}p_{112}p_{332} - p_{021}p_{123}p_{302}p_{112}p_{232} \\
& + p_{021}p_{123}p_{302}p_{132}p_{212} + p_{021}p_{133}p_{102}p_{212}p_{322} - p_{021}p_{133}p_{102}p_{222}p_{312} \\
& - p_{021}p_{133}p_{202}p_{112}p_{322} + p_{021}p_{133}p_{202}p_{122}p_{312} + p_{021}p_{133}p_{302}p_{112}p_{222} \\
& - p_{021}p_{133}p_{302}p_{122}p_{212} - p_{023}p_{121}p_{302}p_{132}p_{212} - p_{023}p_{131}p_{102}p_{212}p_{322} \\
& + p_{023}p_{131}p_{102}p_{222}p_{312} + p_{023}p_{131}p_{202}p_{112}p_{322} - p_{023}p_{131}p_{202}p_{122}p_{312} \\
& - p_{023}p_{131}p_{302}p_{112}p_{222} + p_{023}p_{131}p_{302}p_{122}p_{212} + p_{223}p_{121}p_{002}p_{112}p_{332} \\
& - p_{223}p_{121}p_{002}p_{132}p_{312} - p_{223}p_{121}p_{102}p_{012}p_{332} + p_{223}p_{121}p_{102}p_{032}p_{312} \\
& + p_{223}p_{121}p_{302}p_{012}p_{132} - p_{223}p_{121}p_{302}p_{032}p_{112} - p_{223}p_{131}p_{002}p_{112}p_{322} \\
& + p_{223}p_{131}p_{002}p_{122}p_{312} + p_{223}p_{131}p_{102}p_{012}p_{322} - p_{223}p_{131}p_{102}p_{022}p_{312} \\
& - p_{223}p_{131}p_{302}p_{012}p_{122} + p_{223}p_{131}p_{302}p_{022}p_{112} - p_{323}p_{101}p_{012}p_{122}p_{232} \\
& + p_{323}p_{101}p_{012}p_{132}p_{222} + p_{323}p_{101}p_{112}p_{022}p_{232} - p_{323}p_{101}p_{112}p_{032}p_{222} \\
& + p_{121}p_{103}p_{012}p_{222}p_{332} - p_{121}p_{103}p_{212}p_{022}p_{332} - p_{121}p_{103}p_{312}p_{032}p_{222} \\
& - p_{121}p_{113}p_{002}p_{222}p_{332} + p_{121}p_{113}p_{002}p_{232}p_{322} + p_{121}p_{113}p_{202}p_{022}p_{332} \\
& - p_{121}p_{113}p_{202}p_{032}p_{322} - p_{121}p_{113}p_{302}p_{022}p_{232} + p_{121}p_{113}p_{302}p_{032}p_{222}
\end{aligned}$$

To understand base substitution models in this framework, we confront a basic problem:

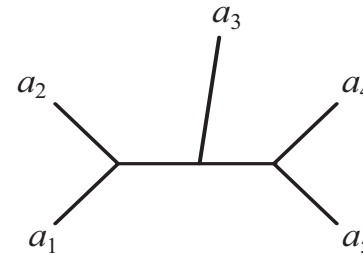
For any fixed tree T and κ , find invariants.

- Ad hoc methods
- Gröbner basis methods on small trees, for simple models
 - compute elimination ideals
 - form of computed invariants often unrevealing (not tied to topology of tree, or other concrete information)
- Complete theoretical analysis...

A representative example:

GM model, 2 states (R/Y, 0/1)

All invariants are known here and tied intimately to branching structure of tree T .



Example: For 2-state GM, 5 taxa

The joint distribution tensor P is $2 \times 2 \times 2 \times 2 \times 2$.

P has two natural *flattenings* according to *splits* in the tree:

$$\{\{a_1, a_2\}, \{a_3, a_4, a_5\}\}, \text{ and } \{\{a_1, a_2, a_3\}, \{a_4, a_5\}\}.$$

The corresponding flattenings are

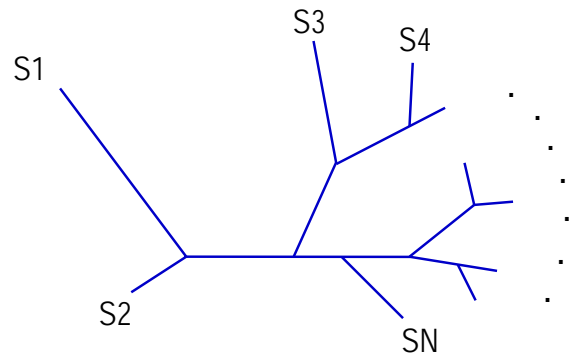
$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}$$

and

$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}.$$

Theorem: For this 5-leaf tree, I_T is generated by all 3×3 minors of these two matrices. (That is, these matrices have rank ≤ 2 .)

This result extends to arbitrary binary trees



Proof uses ideas from algebraic geometry and representation theory.

Similar result are known for GM model, $\kappa > 2$, and ‘group-based’ models

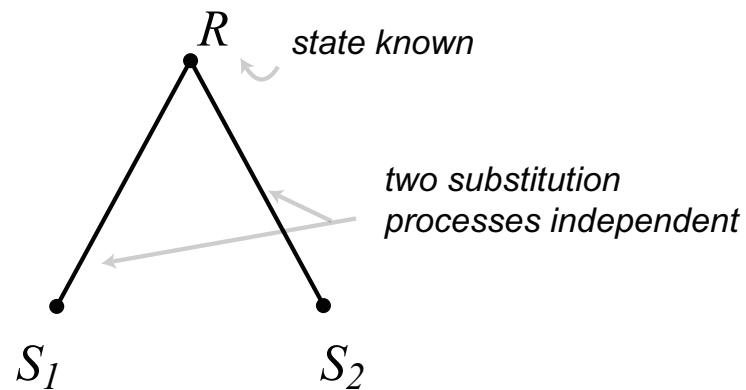
but precise statements are more technical.

(‘Small’ technical gap for GM $\kappa \geq 4$)

What do these polynomials 'say'?

$$3 \times 3 \text{ minors} = 0 \iff \text{rank} = 2$$

reflects modeling assumption of *conditional independence*



Even the analysis of the 3-taxon model has many connections to other areas outside of phylogenetics

- **Statistics:** Latent class models
- **Multilinear algebra:** Tensor rank
- **Algebraic geometry:** Higher secant varieties of Segre varieties,
$$\text{Sec}^k(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \mathbb{P}^{k-1})$$
- **Computational complexity:** Optimal matrix multiplication
- **Quantum Mechanics:** Entanglement
- **Engineering:** Signal decomposition

III. More realistic models

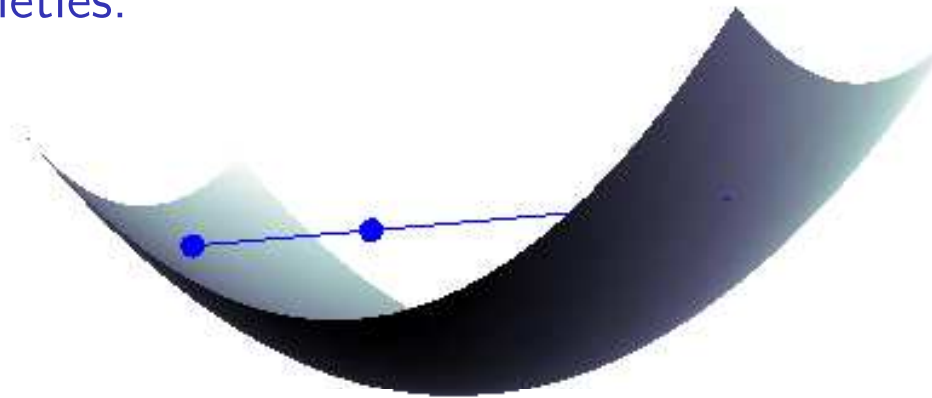
More sophisticated models of the base substitution process can incorporate more biological realism with goal of improving tree inference.

Mixture models

- Allow for a finite number of classes of sites (slow, medium, fast evolving).
- Each class has its own set of numerical parameters, but all share the same tree parameter.
- Additional parameters give sizes of classes.

Two ways of viewing:

1) Secant varieties:



$$P = \delta P_{fast} + (1 - \delta) P_{slow}$$

2) Block models:

On internal edge of tree:

$$M_e = \begin{pmatrix} M_{fast} & 0 \\ 0 & M_{slow} \end{pmatrix}$$

8×8 for DNA

On terminal branches:

$$M_e = \begin{pmatrix} M_{fast} & 0 \\ 0 & M_{slow} \end{pmatrix} \begin{pmatrix} I \\ I \end{pmatrix} = \begin{pmatrix} M_{fast} \\ M_{slow} \end{pmatrix}$$

8×4 for DNA

Theorem: The tree parameter is identifiable for generic parameters for DNA mixture models with 3 or fewer classes.

Proof: Three steps –

- 1) Construct phylogenetic invariants for algebraic model.
- 2) Show those invariants distinguish between trees.
- 3) Specialize to continuous-time models

Note: *Essentially nothing* has been proved about numerical parameters for mixture models. From simulations, it is suspected that they are identifiable for generic parameters.

Continuous-time models are usually used for data analysis

Q an instantaneous rate matrix, fixed over tree

$M = \exp(Qt)$, where t is an edge length (elapsed time)

- sometimes plausible model
- ‘small’ number of parameters (statistically, computationally desirable)
- necessary to infer time of speciation
- can allow instantaneous switching between rate classes in mixture models (covarion models)

Such models embed in algebraic ones, and most of our theoretical understanding of them is from specializing algebraic results.

Ex: Covarion models

(block models with 'switching' between classes)

$M = \exp(Qt)$, with

$$Q = \left(\begin{array}{c|c} \tilde{Q}_1 - s_1 I & s_1 I \\ \hline & \\ s_2 I & \tilde{Q}_2 - s_2 I \end{array} \right)$$

Tree and numerical parameters are identifiable.

Model is **inherently non-algebraic**,

... yet proof of identifiability depends on algebraic approach

Final points:

- Models of sequence evolution are still developing
- As models become more elaborate, we need to understand identifiability (the limits of valid inference)
- Algebraic view of phylogenetic models is both natural and powerful
- These models have rich algebraic structure, provide enriching examples/problems for algebraic geometry
- Different perspectives are valuable.