Progress and potential for phylogenetic invariants

Elizabeth S. Allman

Department of Mathematics and Statistics
University of Southern Maine / University of Alaska

Mathematical Biosciences Institute
Columbus, Ohio   September 26, 2005

Joint work with

John A. Rhodes

Department of Mathematics and Statistics

University of Alaska

For phylogenetic inference,

the data are observed pattern frequencies in aligned sequences:

```
Strepsiptera   AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGAT...
Aedes          AGGCTCAGTATAACACTATAATTTACAAGATCATTGGAT...
Drosophila     AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGAT...
Flea           TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGAT...
Meloe          AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...
Tenebrio       AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGAT...
```

$$\widehat{p}_{AAAAAA} = \frac{\textit{\# observations of } AAAAAA}{\textit{sequence length}}, \text{ etc.}$$

which, assuming a model of molecular evolution along a tree, are estimators for the expected pattern frequencies p_{AAAAAA}, etc.

Phylogenetic Invariants are polynomials in variables $p_{A\ldots A}$ that vanish on expected pattern frequencies.

For example, the stochastic invariant

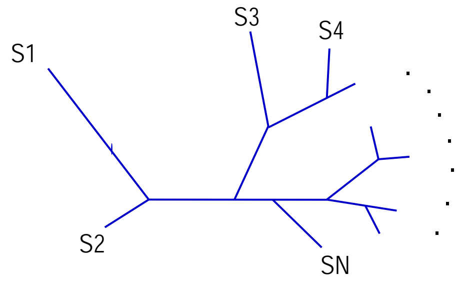
$$\sum_{\mathit{all\ \ patterns}} p_{\mathit{pattern}} = 1$$

Other invariants are typically higher degree and reflect both

- the choice of mutation model and
- the topological branching structure of the tree.

Modeling molecular mutation 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)

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

Model the base substitution process at a single site

for multiple sites, assume i. i. d.

Some models:

general Markov model (GM):

arbitrary root distribution
arbitrary Markov matrices M_e on each edge

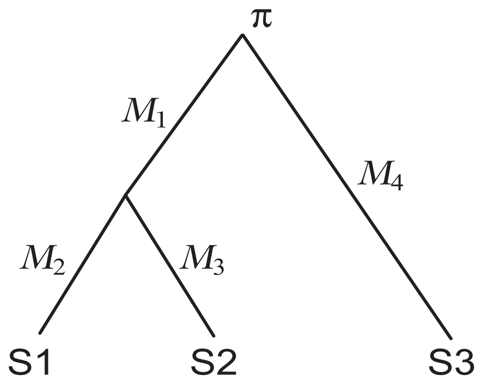
group-based models (JC, K3ST):

equal base frequencies
special form of Markov matrices M_e

mixture models (GM+I, GTR+I, etc.): – later

Given a model,

for example, GM with parameters T, $\boldsymbol{\pi}$,$\{M_e\}$,

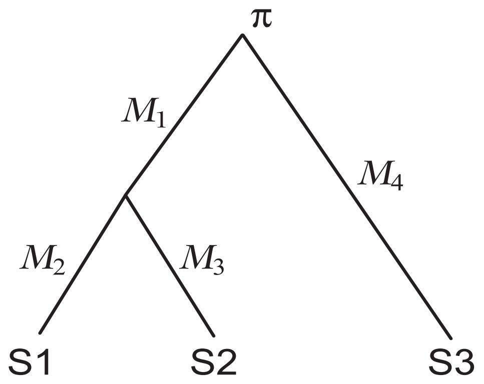


compute the expected pattern frequencies p_{ijk}.

These are gathered in the (table / array / tensor), the

joint distribution P of bases at leaves

$$P(i,j,k) = p_{ijk} = P(S1 = i, S2 = j, S3 = k)$$



E.g.

$$p_{ijk} = \sum_{l=1}^{4} \sum_{m=1}^{4} \pi_l M_1(l,m) M_2(m,i) M_3(m,j) M_4(l,k)$$

$P = (p_{ijk})$ is a $4 \times 4 \times 4$ tensor

with entries that

are polynomial in the stochastic parameters

can be estimated by $\widehat{p}_{ijk}$.

For a fixed tree T,

the joint distribution map defines a polynomial parameterization:

Parameter Space $\longrightarrow$ Joint Distribution space

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

Number N of stochastic parameters is large:

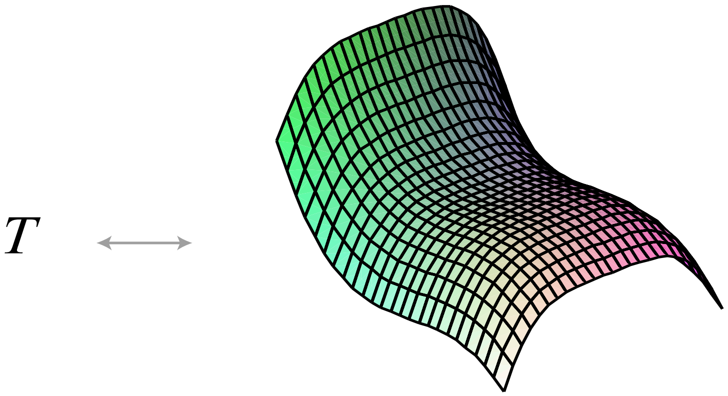
(GM: $N = (\kappa - 1) + (2n-3)\kappa(\kappa-1)$)

Joint distribution $P = (p_{i_1 \dots i_n})$ is large: $\underbrace{\kappa \times \cdots \times \kappa}_{n}$

Extending this polynomial map to the complex setting, then for any model ϕ_T defines a higher-dimensional surface, an algebraic variety.

$$\phi_T : \mathbb{C}^N \longrightarrow \mathbb{C}^{\kappa^n}$$

The closure of the image $\overline{\phi_T(\mathbb{C}^N)}$ is the *phylogenetic variety* V_T.



This associates each tree T to a parameterized surface V_T.

Note:

Extension to $\mathbb{C}$ is appropriate setting for studying polynomial maps.

Restriction to stochastic parameters $\in [0,1]$ – work for the future.

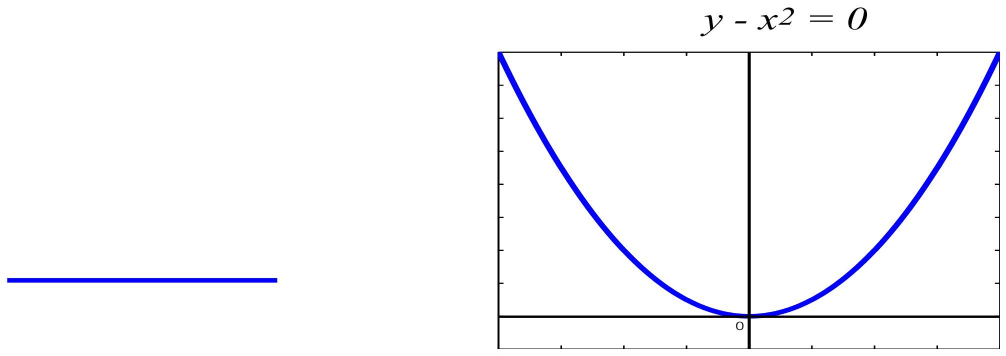
$$\phi_T : \mathbb{C}^N \longrightarrow \mathbb{C}^{\kappa^n}$$

By general principles of algebraic geometry, the phylogenetic variety is also the zero set of a collection of polynomials, the

phylogenetic ideal I_T.

The polynomials in the phylogenetic ideal are *precisely* the phylogenetic invariants for T.

Eg.



$$\phi_T : \quad t \quad \mapsto \quad (t, t^2)$$

Therefore, $I_T = (y - x^2)$.

(Warning: How to find polynomials in I_T is not obvious.)

Example: Consider the GM model and a particular tree T.

For each choice of numerical parameters $s = (\boldsymbol{\pi}, \{M_e\})$, we get a joint distribution $P_0 = \phi_T(s)$, a point on V_T.

And, for each polynomial $f \in I_T$,

$$f(P_0) = 0$$

The vanishing of phylogenetic invariants indicates that a joint distribution P_0 arises from model parameters for GM (both discrete and numerical, though possibly complex).

Using invariants for inference ...

In 1987,

Cavender and Felsenstein

Lake

proposed using phylogenetic invariants for tree inference.

Idea was to evaluate invariants at observed pattern frequencies in aligned sequences (data):

a: ATTAGGTACATGATTAG
b: ATTCGGTACATGATTAG
c: ATTCGCTACATGATCCG
d: ATTTGCTACATGTTCCG

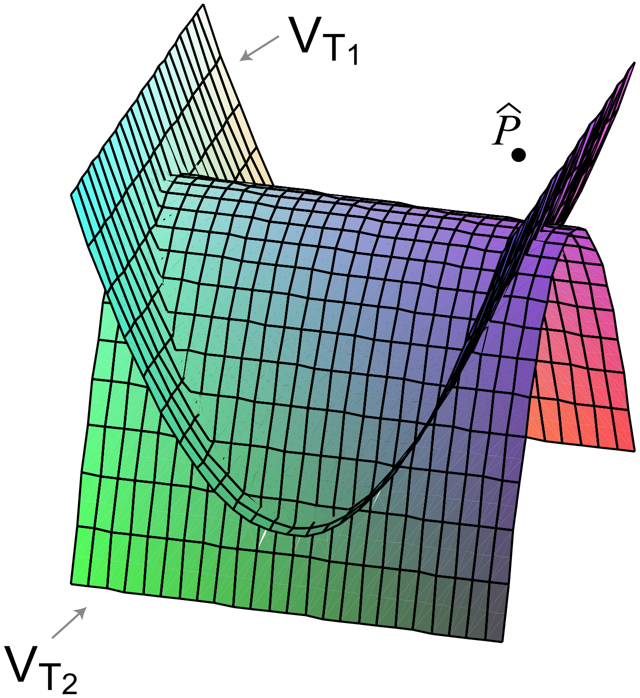
$$\widehat{p}_{AAAA} = 3/17,\ \widehat{p}_{ACCT} = 1/17, \ldots$$

If T, M, are the correct tree and mutation model relating the sequences, then $\widehat{P} = (\widehat{p}_{ijkl})$ and $\widehat{P} \approx P_0$.

Since $f(P_0) = 0$, then $f(\widehat{P}) \approx 0$

‘Plan’ of Implementation: Given data $\widehat{P}$,

- Fix M.
- For each T,
  - Find some/most/all invariants f for V_T
  - Test if $f(\widehat{P}) \approx 0$.
- Return tree for which $\widehat{P}$ “ $\in$ ” V_T (as best possible)



T_1 versus T_2?

Comments:

This method is statistically consistent.

More generators of I_T known $\rightsquigarrow$ improved tree inference.

Decouples tree inference from parameter estimation (though a choice of model is still necessary).

In general, invariants have been difficult to find.

best understanding is for group-based models, GM, ...

Invariants may have other uses.....

Drawbacks:

Invariants will not be identically zero, only close to zero

- statistical issues (finite length sequences, imperfect model)
- algebraic issues (evaluation at points off V_T,
  precise form affects “near” vanishing)

Simulation study: Lake’s invariants displayed poor performance.

- Looking for ‘surface’ V_T in linear subspace which contains it

Theoretically, techniques from computational algebra (Gröbner bases) could find all invariants, but

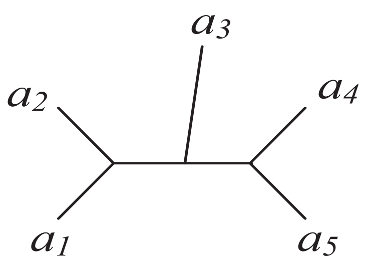
- only for very small trees and simple models
- form of invariants coming from such computational methods is unrevealing (not tied to topology of tree, or other concrete information)

Phylogenetic invariants for the general Markov model

Simplest case:

$\kappa = 2$ states (R/Y)

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



Example: For GM, $\kappa = 2$,

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

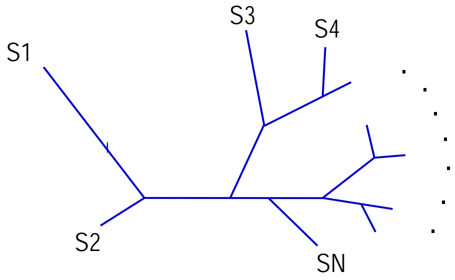
$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111}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

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.)

Theorem: For $\kappa = 2$, any bifurcating T, GM, the phylogenetic ideal I_T is generated by *edge invariants* (3×3 determinants associated to flattenings on edges).



(This was conjectured by Pachter-Sturmfels)

Implications: Via rank of flattenings, we can potentially say something about data's support for a *particular edge* of a phylogenetic tree.

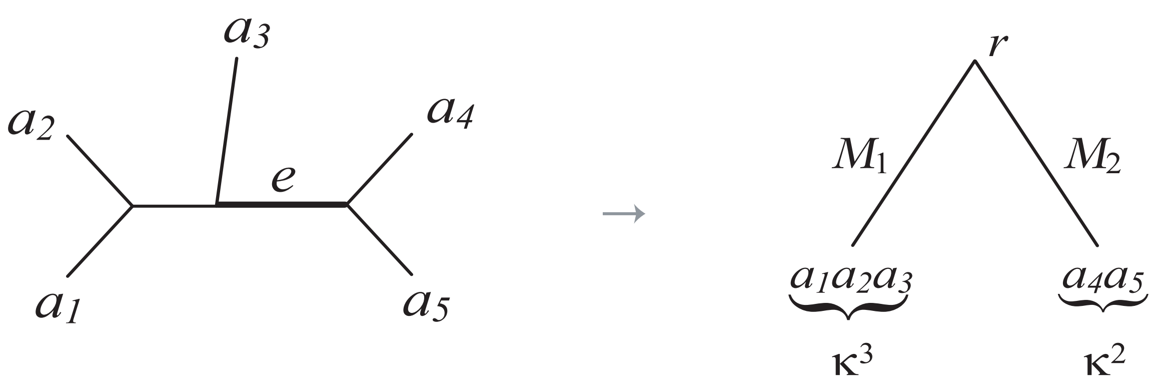
Furthermore,

$$\text{support for all edges} = \text{support for tree}$$

(when $\kappa = 2$).

A glimpse at the proof...

Related to an edge e in T is a 'simpler' graphical model:



a_3
a_2
a_4
e
a_1
a_5
→
r
M_1
M_2
$\underbrace{a_1 a_2 a_3}_{\kappa^3}$
$\underbrace{a_4 a_5}_{\kappa^2}$

for

M_1, a $\kappa \times \kappa^3$ matrix

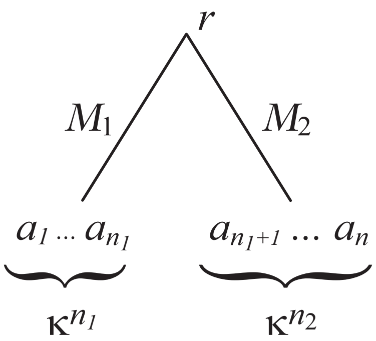
M_2, a $\kappa \times \kappa^2$ matrix

5-dim $\kappa \times \cdots \times \kappa$ tensor $P \;\longrightarrow\; \kappa^3 \times \kappa^2$ matrix $\mathrm{Flat}_e(P)$

For an n-taxon tree,

with internal edge e inducing the split $a_1 \cdots a_{n_1} \mid a_{n_1+1} \cdots a_n$

n-dim $\kappa \times \cdots \times \kappa$ tensor P $\rightarrow$ $\kappa^{n_1} \times \kappa^{n_2}$ matrix $\mathrm{Flat}_e(P)$



Thus,

$$\mathrm{Flat}_e(P) = M_1^T \,\mathrm{diag}(\pi_r) M_2,$$

and $\mathrm{Flat}_e(P)$ is rank κ matrix (at most).

But ...

Matrix rank is well-understood to be given by a polynomial condition:

$$\mathrm{Flat}_e(P) = M_1^T \,\mathrm{diag}(\pi_r) M_2 \text{ is of rank at most } \kappa \;\longleftrightarrow$$

$$\text{all } (\kappa+1)\times(\kappa+1) \text{ minors vanish.}$$

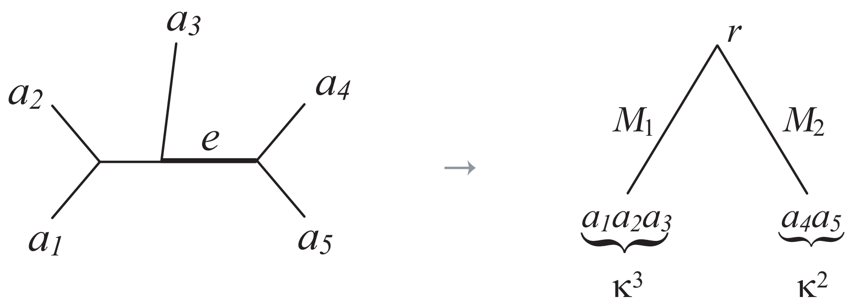
These polynomials are the edge invariants for a tree T.

Proof that this gives all invariants for $\kappa = 2$ states uses ideas from algebraic geometry and representation theory.

DNA: $\kappa = 4$ states

Current results are less complete...

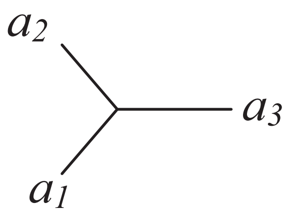
The rank argument establishing the existence of edge invariants works for any number of states κ and any tree relating at least 4 taxa.


a_3
a_2
a_4
e
a_1
a_5
$\rightarrow$
r
M_1
M_2
$\underbrace{a_1a_2a_3}_{\kappa^3}$
$\underbrace{a_4a_5}_{\kappa^2}$

So, edge invariants exist — edge flattenings are rank 4 matrices.

However....

Example: 3 taxa, $\kappa = 4$



Joint distribution $P = (p_{ijk})$, a $4 \times 4 \times 4$ tensor of rank 4

There are no edge invariants, but ....

Theorem: A 1728-dim space of all quintic polynomials in I_T can be explicitly constructed. (i.e. all phylogenetic invariants of degree 5)

For instance,

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

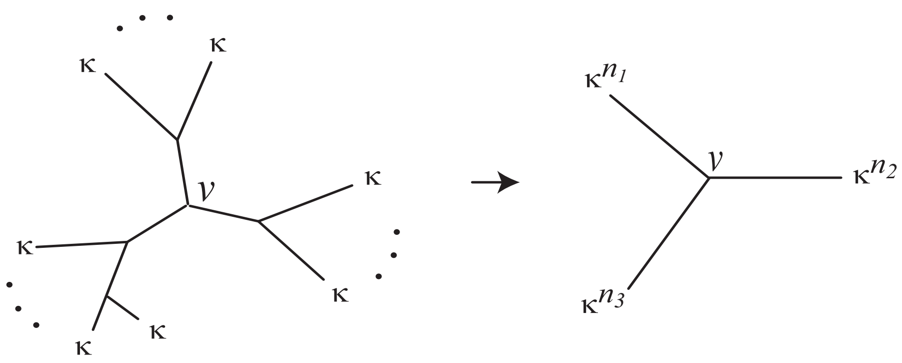
$$
\begin{array}{l}
+p_{323}p_{121}p_{002}p_{132}p_{212} + p_{323}p_{121}p_{102}p_{012}p_{232} - p_{323}p_{121}p_{102}p_{032}p_{212} \\
-p_{323}p_{121}p_{202}p_{012}p_{132} + p_{323}p_{121}p_{202}p_{032}p_{112} + p_{323}p_{131}p_{002}p_{112}p_{222} \\
-p_{323}p_{131}p_{002}p_{122}p_{212} - p_{323}p_{131}p_{102}p_{012}p_{222} + 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}
\end{array}
$$

$$
\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}
$$

Comments:

- For $\kappa > 2$ states, edge invariants can not possibly be enough.
- This formula seems vaguely 'determinantal', and in fact is — there are better ways to evaluate it than this formula.
- The quintics *locally* define V_T where it matters most for biology.

For an arbitrary tree, focus on a node:



For $P \in V_T$, flatten

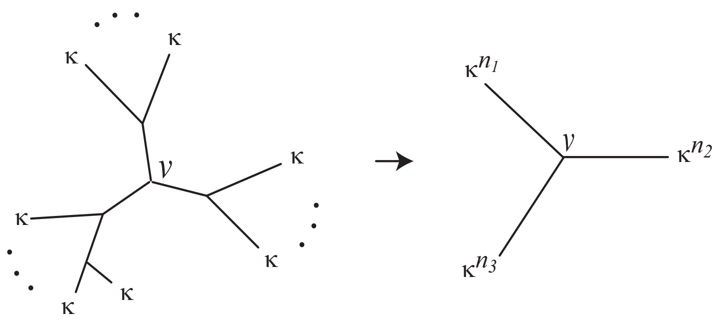
$$P \mapsto \mathrm{Flat}_v(P),$$

a $4^{n_1} \times 4^{n_2} \times 4^{n_3}$ tensor, $n_1 + n_2 + n_3 = n$.

Then

$\mathrm{Flat}_v(P)$ is a 3-dimensional tensor of rank 4

"$\mathrm{Flat}_v(P)$ is a 3-dimensional tensor of rank 4"



- $\mathrm{Flat}_v(P)$ is the sum of four 3-dim. tensors, one for each possible state – A, C, G, T – at the internal node, i.e.

$$\mathrm{Flat}_v(P) = (p_{ijk})_A + (p_{ijk})_C + (p_{ijk})_G + (p_{ijk})_T$$

- Using tensor rank to understand joint distributions P is very natural – basic idea is independence of substitutions on various edges, conditioned on state at the internal node.

Main result for $\kappa > 2$:

Theorem: For any κ, given all invariants associated to the 3-taxon tree, we can explicitly construct defining polynomials for V_T for GM model on any trivalent tree T.

In the case of DNA, $(\kappa = 4)$, one obstacle remains:

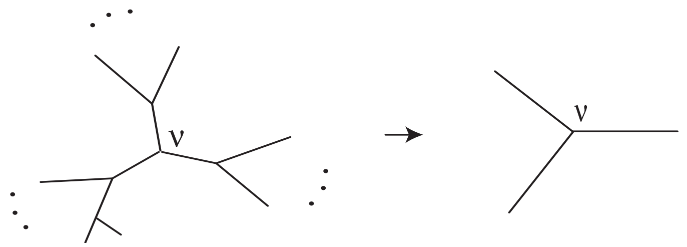
> determining an explicit set of polynomials defining V_{T_3}

However, this is a technical problem in algebraic geometry, not phylogenetics (1728-dim. space, local definition, saturation, ...).

Informally ...

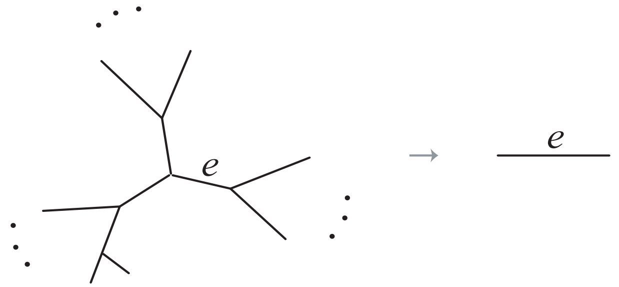
the local structure of a tree determines a collection of phylogenetic invariants defining the variety.

Flatten for all vertices v:



n-dim $\kappa \times \cdots \times \kappa$ tensor $P \;\longrightarrow\;$ 3-dim $\kappa^{n_1} \times \kappa^{n_2} \times \kappa^{n_3}$ tensor

and for all edges e:



n-dim $\kappa \times \cdots \times \kappa$ tensor $P \;\longrightarrow\;$ $\kappa^{n_1} \times \kappa^{n_2}$ matrix

Implications:

Local structure of invariants may be used to test for one tree feature at a time without considering all the details.

More specifically, via rank of flattenings, we can potentially say something about data's support for a *particular edge* or *particular node* of a phylogenetic tree. Furthermore,

$$\text{support for all edges and nodes} = \text{support for tree}$$

Beyond finding invariants ...

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

- Algebraic mixture models: JC+I, GM+I, GM+GM+GM, etc.
- Continuous models: GTR+I+Γ
- Covarion models

The viewpoint of algebraic geometry has led to new insight into some familiar problems in model-based phylogenetics.

- Identifiability results
- Parameter Estimation
- Maximum Likelihood solutions
- Tree Construction algorithms (preliminary)

Question: Given a joint distribution of bases at the leaves from some model, are model parameters (T, stochastic) *identifiable*?

Question is usually addressed sequentially:

Assuming P rises from model parameters,

can you recover the tree parameter T?

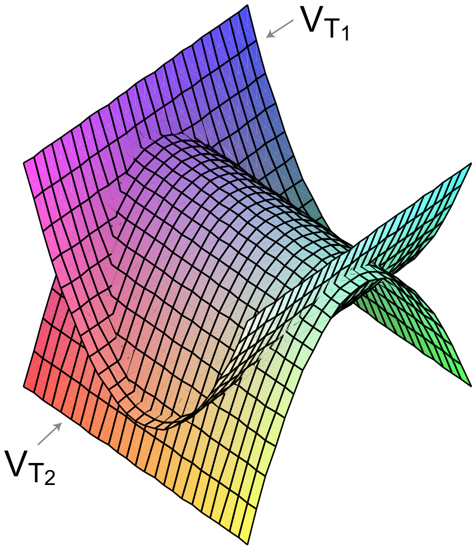
can you recover the stochastic parameters?

(and how many taxa at a time must be considered to do this?)

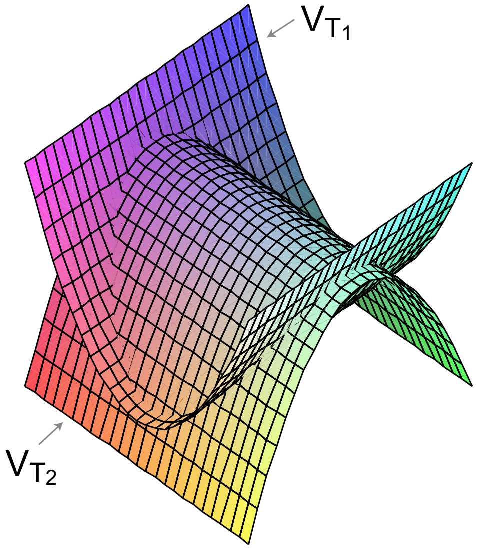
Identifiability is needed to prove the consistency of inference methods such as Maximum Likelihood.

Identifiability of Tree topology:

Since V_{T_1} and V_{T_2} are irreducible varieties of the same dimension, then we ask



Are the two varieties distinct?



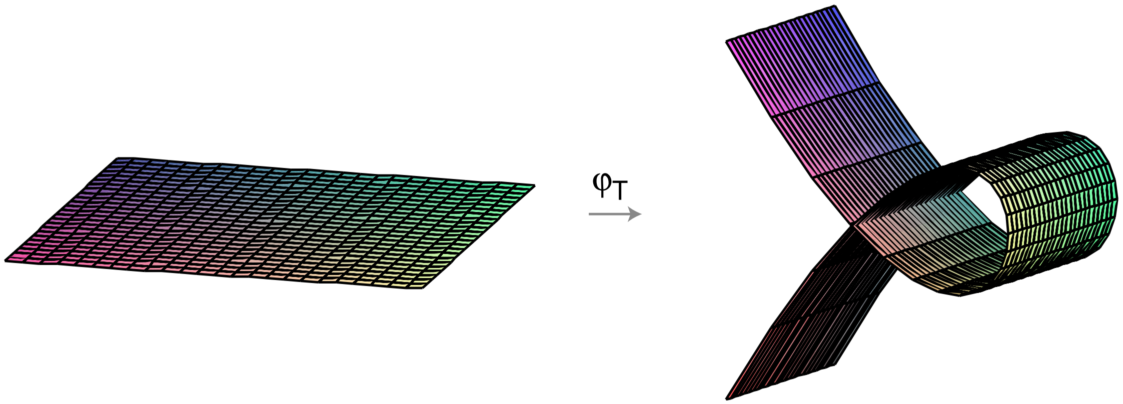
If $V_{T_1} = V_{T_2}$, then tree identifiability fails for all parameter choices.

Otherwise, $V_{T_1} \cap V_{T_2}$ is of lower dimension and the tree is identifiable for *generic* parameters.

Stochastic parameters:

Identifiability $\equiv$ inverting ϕ_T

For a 'good' parameterization (generically 1-1), all points with non-identifiable parameters lie in singular locus.



Where is V_T non-singular?

(When can you invert ϕ_T?)

New results use invariants to identify T *without* using distances:

Eg. GM+I model

> No known distance formula, yet T is identifiable by vanishing of 4-leaf invariants.

This follows from the observation that the form of invariants associated to edge flattenings for a 4-taxon tree with GM+I model makes it possible to recover the proportion of invariable sites.

Can these be developed into useful estimators of parameters?

Similar invariant-based arguments (re)prove identifiability (tree and stochastic) of other models

- GM,
- GM$+$I,
- GM$+$GM$+\ldots+$GM model ($< \kappa$ summands),
- rates-across-sites models with $< \kappa$ arbitrary rate classes
- Covarion model of Tuffley and Steel (if $\kappa \geq 3$)

- Covarion model of Tuffley-Steel

    8 states at internal nodes

$$A^{\text{on}}, A^{\text{off}}, C^{\text{on}}, C^{\text{off}}, G^{\text{on}}, G^{\text{off}}, T^{\text{on}}, T^{\text{off}}$$

    4 observable states at leaves A, C, G, T

    $M_e = \exp(Qt_e)$ where Q is 8×8 rate matrix of special form

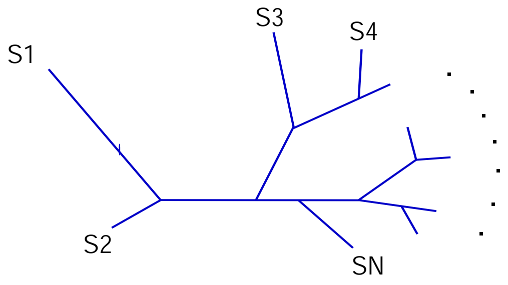
This model allows sites to switch between variable/invariable modes in different parts of tree, increases biological realism.

There is no known distance for this model, but T identifiable through 4-leaf invariants.

Covarion result is first proved for more general model, with purely algebraic definition (no common rate matrix).

Specialization to covarion model involves analytic variety.

Eg.



$M_{internal}$ of size 8×8

$M_{pendant}$ of size 8×4

Compute the joint distribution P.

Internal edge flattening according to correct T has rank ≤ 8, flattening according to wrong tree gives higher rank.

Invariants as Constraints in Maximum Likelihood problems:

Understanding the geometry of likelihood function:

(Chor, Holland, Hendy, Penny, 2000)

Use invariants to show that the likelihood function can have multiple maxima or even a continuum of maxima

– Constrained optimization problem for Neyman model, 4-taxa

Exact ML Optimization: For small trees, use invariants in constrained optimization problem for exact solution of ML problem via computational algebra

(Chor, Khetan, Snir, 2003) Computations for Neyman model.

(Hoşten, Khetan, Sturmfels, 2004) Computations for JC model.

(Chor, Hendy, Snir, 2005) Computations for JC model.

Tree construction using GM invariants/SVD (Eriksson, 2005).

- Focus on edge invariants = matrix rank conditions
- Rank of (noisy) matrix can be computed well via SVD
- Algorithm constructs tree via joining neighbors from outside

to appear, *Algebraic Statistics for Computational Biology*, CUP

References

E. S. Allman and J. A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.

E. S. Allman and J. A. Rhodes. Phylogenetic invariants for stationary base composition. *J. Symbolic Comp.*, 2004. to appear, `arXiv:q-bio.PE/0407035`.

E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. 2004. preprint, `arXiv:math.AG/0410604`.

E. S. Allman and J. A. Rhodes. Phylogenetic invariants and parameter recovery for the general Markov plus invariable sites model. 2005. in preparation.

E. S. Allman and J. A. Rhodes. The identifiability of a general phylogenetic model, with application to the covarion model. 2005. in preparation.

J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. of Class.*, 4:57–71, 1987.

J. T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.

B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Bio. and Evol.*, 17:1529–1541, 2000.

B. Chor, M. D. Hendy, and S. Snir. Maximum likelihood Jukes-Cantor triplets: analytic solutions. 2005. preprint, `arXiv:q-bio.PE/0505054`.

N. Eriksson. Tree construction using singular value decomposition. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005. to appear.

S. N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

S. Hosten, A. Khetan, and B. Sturmfels. Solving the likelihood equations. 2004. preprint, arXiv:math.ST/0408270.

J. P. Huelsenbeck. Performance of phylogenetic methods in simulation. *Sys. Biol.*, 44(1):17–48, 1995.

J. A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Bio. Evol.*, 4:167–191, 1987.

L. Pachter and B. Sturmfels, editors. *Algebraic statistics for computational biology*. Cambridge University Press, 2005. to appear.

Chris Tuffley and Mike Steel. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.*, 147(1):63–91, 1998.

M. A. Steel, L.A. Székely, and M. D. Hendy. Reconstructing trees from sequences whose sites evolve at variable rates. *J. Comput. Biol.*, 1(2):153–163, 1994.

Current contact information:

e.allman@uaf.edu

j.rhodes@uaf.edu