Evolutionary Trees and phylogenetics: An algebraic perspective

Elizabeth S. Allman

University of Alaska
Fairbanks

University of Nevada
Las Vegas

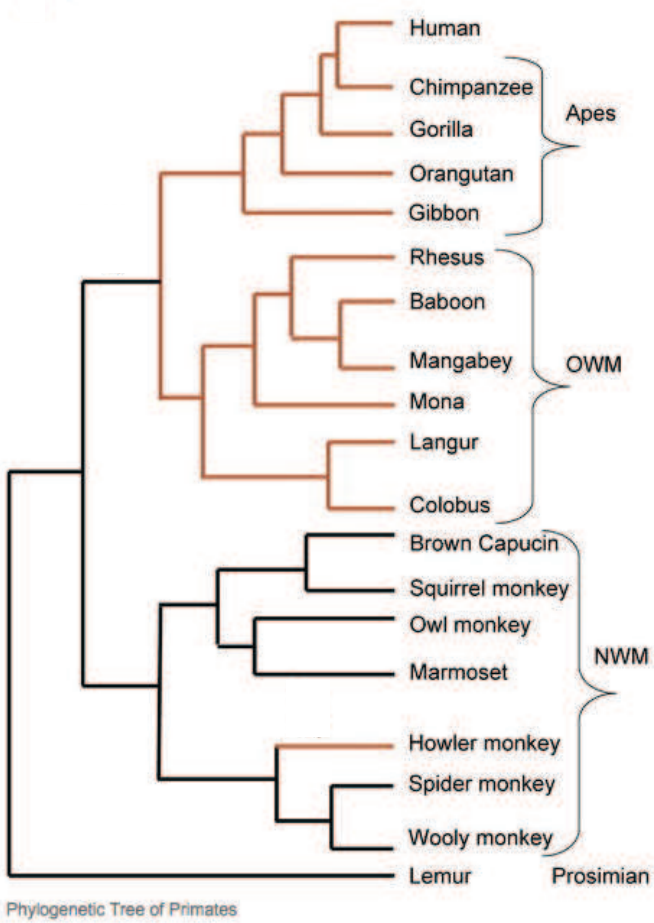
April 30, 2011

2011 AMS Western Section Meeting

Outline

- Biological problem
- Mathematical introduction
- Phylogenetic models
- Contributions from the algebraic geometry viewpoint
- Problems from coalescent theory

Phylogenetics



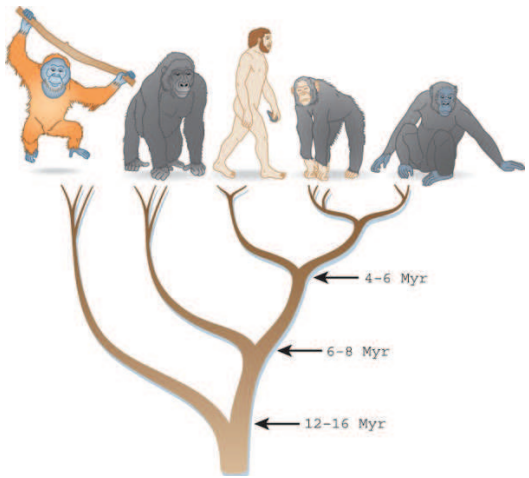
Phylogenetic Tree of Primates

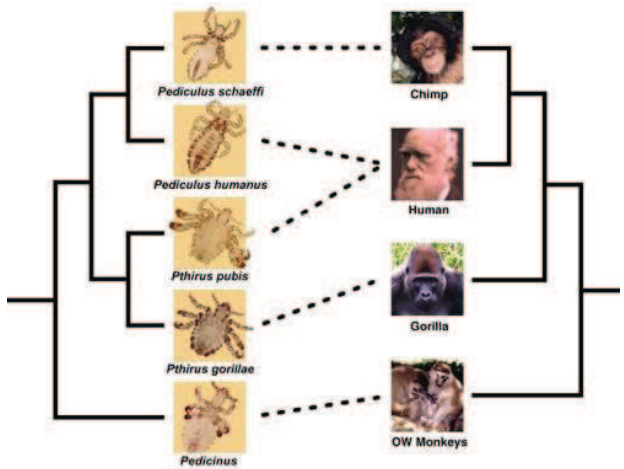
PHYLOGENETICS

is the discipline concerned with inferring evolutionary relationships.

Some compelling questions

- Understanding species relationships
- Biodiversity and conservation
- Dating of speciation events
- Species delimitation
- Deviations from tree-like evolution
- Co-evolution, radiations and bottlenecks, population genetics, ...





Any comparison of organisms (called taxa) should take into account evolutionary correlations.

The data

For a phylogenetic analysis from molecular data, aligned DNA (or protein, etc.) sequences are used:

Primate mitochondrial DNA sequences, HindIII gene

from Hayasaka, K., T. Gojobori, and S. Horai. MBE (1988) 5:626-644.

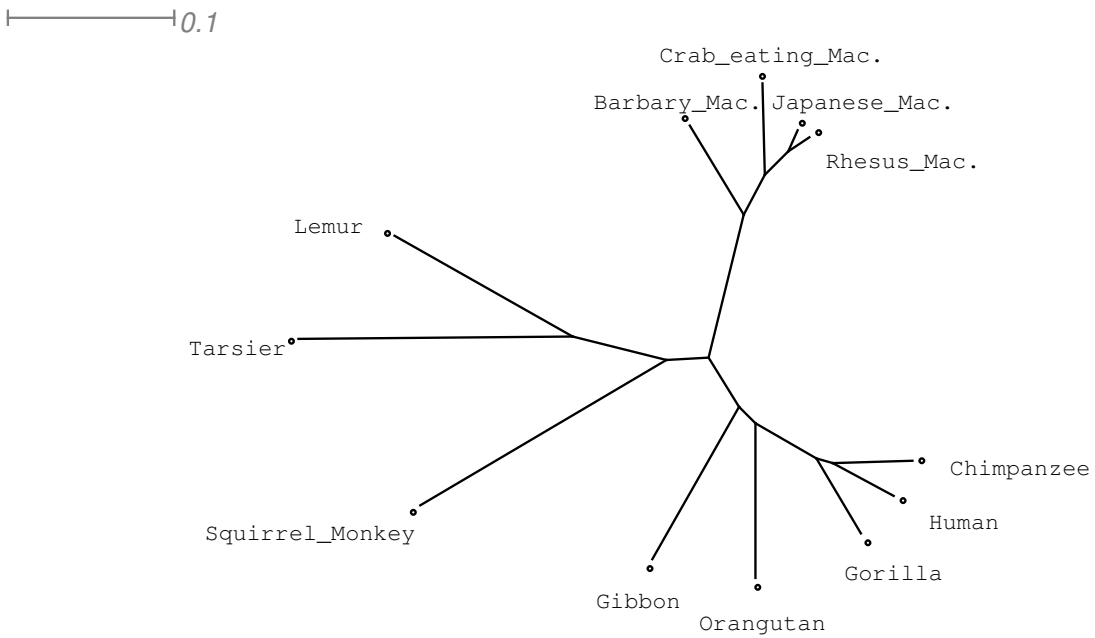
```
Gorilla         AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCAT...
Orangutan       AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCT...
Human           AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCT...
Chimpanzee      AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT...
Gibbon          AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTT...
Crab-eat_Mac    AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTT...
Lemur           AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCAT...
Barbary_Mac     AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTT...
Japanese_Mac    AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTT...
Squirrel_Mon    AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGT...
Rhesus_Mac      AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTT...
Tarsier         AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCT...
```

of length 898 sites, Tree reconstruction (NJ) leads to....

Phylogenetic tree

Primate mitochondrial DNA sequences, HindIII gene

from Hayasaka, K., T. Gojobori, and S. Horai. MBE (1988) 5:626-644.



The data

For phylogenetic inference,
the data are *observed pattern frequencies* in aligned sequences:

```
Gorilla      AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCAT...
Orangutan    AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCT...
Human        AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCT...
Chimpanzee   AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT...
Gibbon       AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTT...
```

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

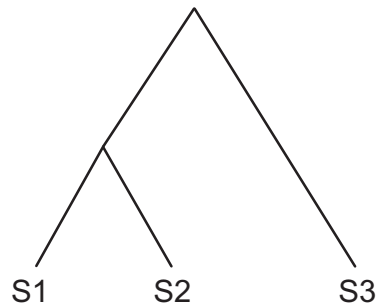
which, assuming a model of molecular evolution along a tree, are estimators for the true joint distribution p_{AAAAA}, etc.

Modeling molecular evolution along a tree T

Fix an n-taxon (binary) rooted, leaf-labelled tree T,

root = most recent common ancestor

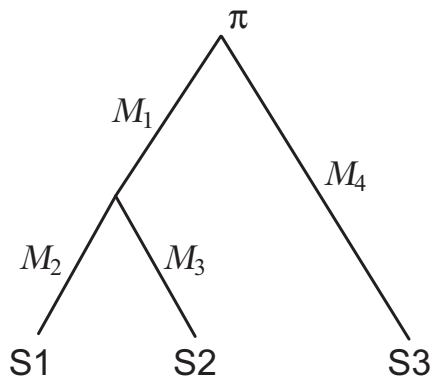
leaves = currently extant taxa



k states at each node,

$k = 4$ (A,C,G,T),
$k = 2$ (R={A,G},Y={C,T}),
$k = 20$ (proteins)
$k = 61$ (codons=triplets of A,C,G,T)

General Markov model on T

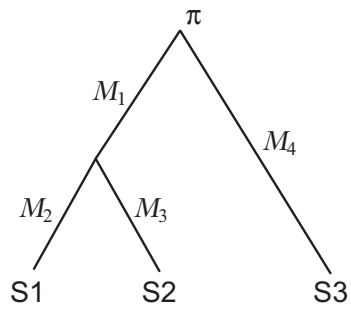


$$\text{Model parameters} = \begin{cases} \text{tree topology} \\ \text{root distribution vector } \boldsymbol{\pi} = (\pi_i) \\ \text{Markov matrix on each edge } M_e \end{cases}$$

Model describes evolution at a single site in sequence.

General Markov model on T

More specifically,



- States $1, 2, \ldots, k$ $\quad (A, C, G, T \rightsquigarrow 1, 2, 3, 4)$
- State at root given by probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$; $\sum \pi_i = 1$.
- On edge e, Markov matrix M_e give probs. of state change,

$$M_e(i, j) = \Pr(j \text{ at end} \mid i \text{ at start})$$

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

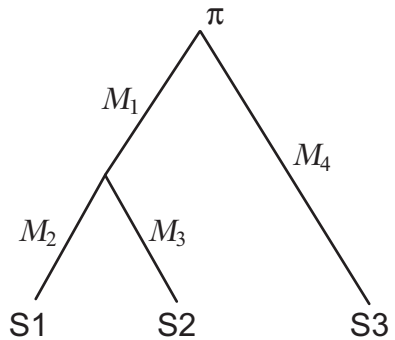
(More restrictive/expansive models are used in practice ....)

GM(k) model

For a fixed topology T, there is map ψ_T taking parameters $\theta = \{\pi, \{M_e\}\}$ to the

joint distribution P of states at leaves

with entries



$$p_{i_1 i_2 i_3} = \sum_{l=1}^{k} \sum_{m=1}^{k} \pi_l M_1(l, m) M_2(m, i_1) M_3(m, i_2) M_4(l, i_3).$$

$P = (p_{i_1 i_2 i_3}) = \psi_T(\theta)$ is a $k \times k \times k$ tensor (table, distribution).

The pattern probabilities $p_{\mathbf{i}}$

- can be estimated from data by $\widehat{p}_{i_1 i_2 i_3}$
- are polynomial in the stochastic parameters
- have polynomial structure determined by the topology of T

Viewpoint of algebraic geometry

A variety is the zero set of a collection (ideal) of polynomials.

Some *varieties* arise from *polynomial parameterization maps* by taking the Zariski closure. Then, $\mathrm{Im}(\psi)$ is a dense subset of V.

$$\psi : \Theta \xrightarrow{\textit{polynomial}} \mathrm{Im}(\psi) \hookrightarrow \overline{\mathrm{Im}(\psi)} = V = \mathcal{Z}(\{f_\alpha\})$$

Viewpoint of algebraic geometry

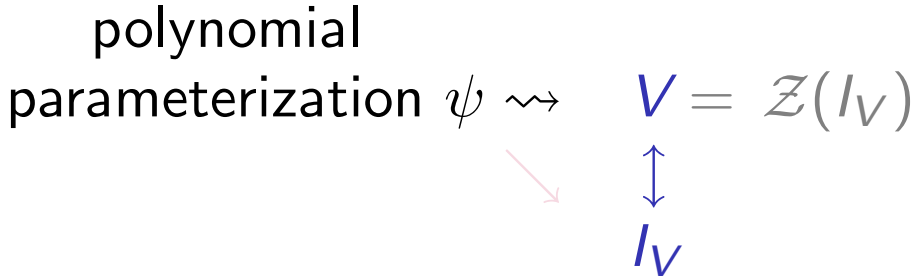
A variety is the zero set of a collection (ideal) of polynomials.

Some *varieties* arise from *polynomial parameterization maps* by taking the Zariski closure. Then, $\mathrm{Im}(\psi)$ is a dense subset of V.

$$\psi : \Theta \xrightarrow{\textit{polynomial}} \mathrm{Im}(\psi) \hookrightarrow \overline{\mathrm{Im}(\psi)} = V = \mathcal{Z}(\{f_\alpha\})$$

Viewpoint of algebraic geometry

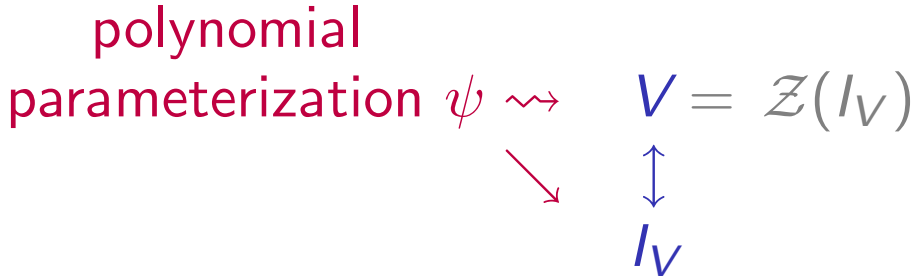
A (geometric) variety V corresponds to its (algebraic) ideal I_V.



The *Implicitization process* passes from the parameterization map ψ and V to the ideal I_V defining V implicitly.

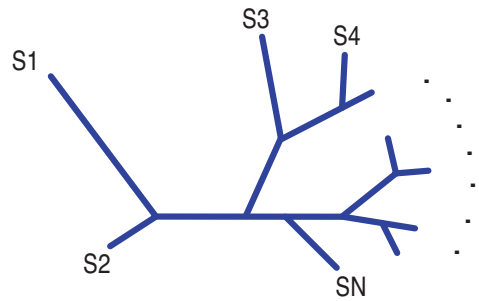
Viewpoint of algebraic geometry

A (geometric) variety V corresponds to its (algebraic) ideal I_V.

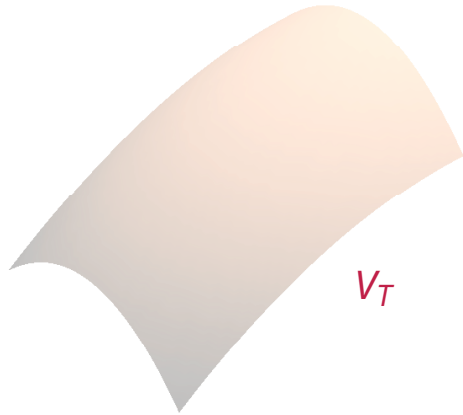


The *Implicitization process* passes from the parameterization map ψ and V to the ideal I_V defining V implicitly.

Viewpoint of algebraic geometry in phylogenetics



$\rightsquigarrow$

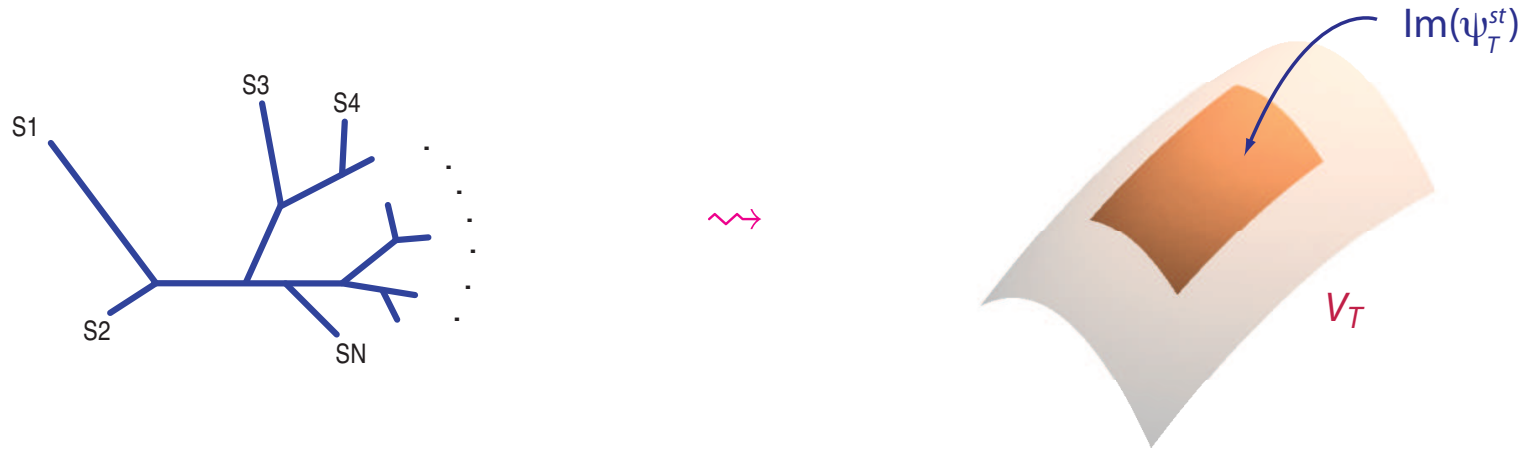


To each tree T, associate a parameterized *phylogenetic variety* V_T, by extending ψ_T to a complex parameter space.

$$T,\ \psi_T \quad \rightsquigarrow \quad \begin{array}{c} V_T \\ \updownarrow \\ I_{V_T} \end{array}$$

The corresponding ideal is the *phylogenetic ideal*.

Viewpoint of algebraic geometry in phylogenetics



Ideally, it would be possible to distinguish the image of stochastic parameters ('the points of interest') from other points on V_T.

If $\theta_T = \{\pi, \{M_e\}\}$ are stochastic parameters on T, then the distribution arising from θ_T is a point P of (biological) interest.

What is the image of

$$\psi_T^{st} : \Theta^{st} \longrightarrow \{P\} \ ?$$

Phylogenetic variety

In summary,

to each tree T, we associate a *phylogenetic variety*.

$$\begin{array}{ccll} T & \rightsquigarrow & V_T & \textit{phylogenetic variety} \\ & & \updownarrow & \\ & & I_T & \textit{phylogenetic ideal} \end{array}$$

V_T can, in principle, be described implicitly by polynomial equations in I_T.

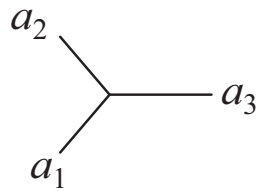
The phylogenetic variety V_T is

- parameterized, and therefore irreducible
- typically of dimension expected; that is, equal to the number of numerical parameters (but with singularities)
- contains stochastic points and lots of 'other' points

Eg. A simple case connected to well-studied objects

For GM(4) on a 3-taxon tree T rooted at central node, (the case of DNA), the pattern probabilities are

$$p_{ijk} = \sum_{f=1}^{4} \pi_f M_1(f,i) M_2(f,j) M_3(f,k),$$



and

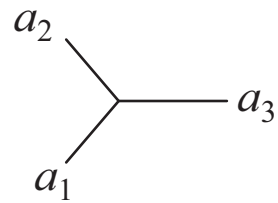
$$P = (p_{ijk}) = p^1_{ijk} + p^2_{ijk} + p^3_{ijk} + p^4_{ijk} \in \mathrm{Im}(\psi_T)$$

is a $4 \times 4 \times 4$ tensor of rank 4.

(Sum of four rank 1 tensors.)

(*tensor rank*: analogy with matrix rank)

Eg. A simple case connected to well-studied objects



For GM(4) on the 3-taxon tree,

$$P \in \mathrm{Im}(\psi_T) = p^1_{ijk} + p^2_{ijk} + p^3_{ijk} + p^4_{ijk}$$

is a rank 4 tensor.

Connections:

- Extended to the projective setting, this V_T is a secant variety
$$\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3).$$
- Applied multilinear algebra: tensor decomposition

A few results: Overview

In what ways has an algebraic perspective helped?

Or spurred new research questions?

I. Phylogenetic invariants (What is the ideal of V_T?)

II. Identifiability results (Is $V_{T_1} = V_{T_2}$?)

III. Understanding when a distribution from a single tree can match a distribution from a mixture of two trees.
(primary decomposition)

IV. From gene trees to species trees.

I. Phylogenetic invariants

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

Traditionally,

$f \in I_T$ is called a *phylogenetic invariant*.

Phylogenetic invariants are polynomials in the pattern probabilities $p_{\mathbf{i}}$, for instance,

$$\sum_{\mathbf{i}} p_{\mathbf{i}} - 1 \in I_T.$$

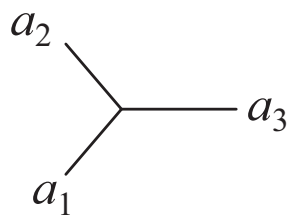
The structure of *informative* invariants depends on the topology of T and choice of substitution model.

I. Phylogenetic invariants

In principle, invariants can be computed

— Gröbner bases, elimination;

In practice, usually not.



Example: 3 taxa, GM(4)

- Trivial invariant: $\sum p_{ijk} - 1$
- There are no homogeneous invariants of degree < 5.
- Full ideal I_T of invariants is unknown.
- A 1728-dim space of *all* quintics in I_T can be explicitly constructed. For instance ...

Eg. Invariant for 3-taxon T

$$
\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} \\
& +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}
\end{aligned}
$$

Eg. Invariant for 3-taxon T

$$
\begin{aligned}
&- 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} \\
&+ 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}
\end{aligned}
$$

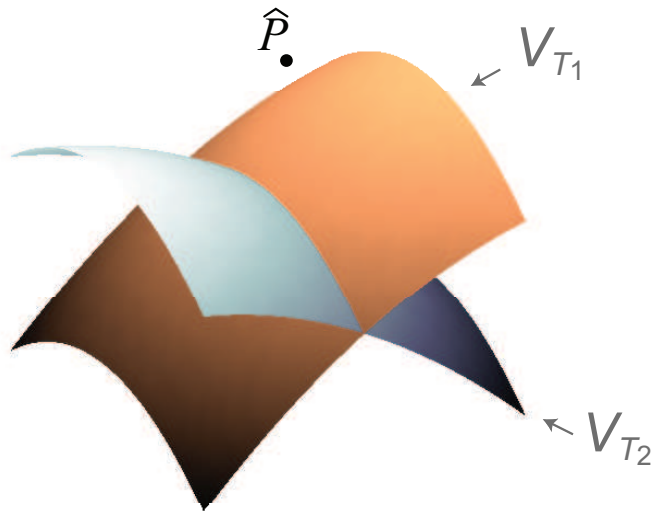
Eg. Invariant for 3-taxon T

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

I. Phylogenetic invariants

Cavender and Felsenstein; Lake 1987

Invariants were originally introduced for inference ....
Idea is to infer the topology first.



If $f(\widehat{P}) \approx 0$ for all $f \in I_T$, then infer data comes from tree T.

I. Phylogenetic invariants

Many researchers have studied phylogenetic invariants

Bates, Buczynska, Casanellas, Cavender, Chifman, Erdös, Eriksson, Evans, Felsenstein, Fernandez-Sanchez, Friedland, Fu, Gross, Hagedorn, Hendy, Jarvis, Lake, Oeding, Petrović, Rhodes, Sankoff, Speed, Steel, Székely, Sturmfels, Sullivant, Sumner, Wisniewski, ...

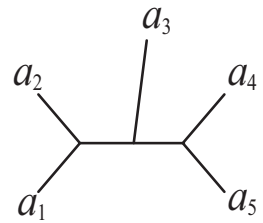
Today's example:

some invariants for GM(k) model,

and their relation to *local features* of T.

Example

For GM(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:

$a_1a_2 \mid a_3a_4a_5$, and $a_1a_2a_3 \mid a_4a_5$.

The corresponding flattenings are

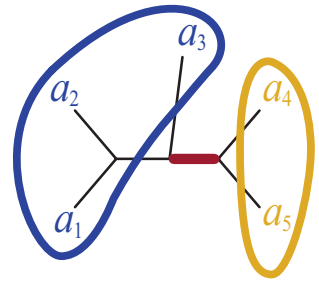
$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111}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

Rows and columns are indexed by states at conglomerate variables $a_1a_2a_3$ and a_4a_5.

and a 4×8 matrix defined analogously.

Example

For GM(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:

$a_1a_2 \mid a_3a_4a_5$, and $\mathbf{a_1a_2a_3} \mid \mathbf{a_4a_5}$.

The corresponding flattenings are

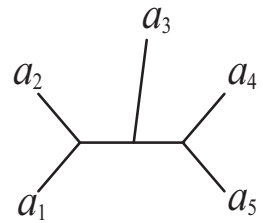
$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111}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

Rows and columns are indexed by states at conglomerate variables $a_1a_2a_3$ and a_4a_5.

and a 4×8 matrix defined analogously.

Example

For GM(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:

$$a_1a_2 \mid a_3a_4a_5, \text{ and } a_1a_2a_3 \mid a_4a_5.$$

The two flattenings are

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

We prove these matrices have rank ≤ 2.

Thus, for this T, I_T contains all 3×3 minors of these two matrices.

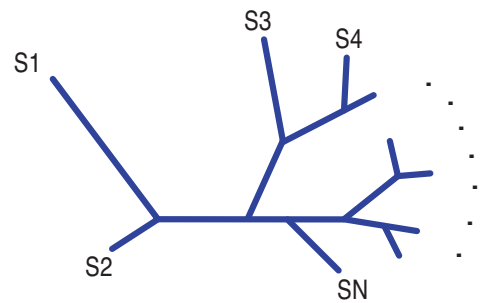
These polynomials are called the edge invariants.

Ideal generation GM(2)

Theorem (A, Rhodes 2008):

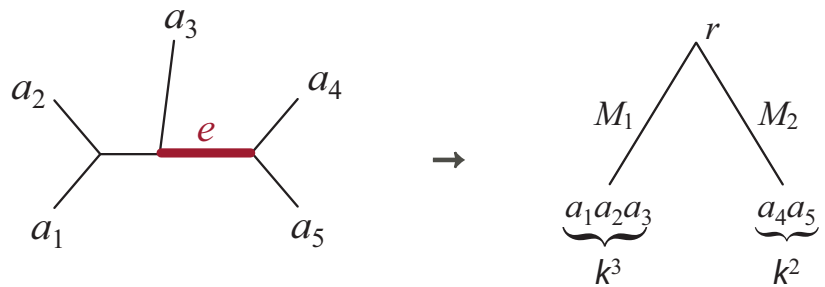
For $k = 2$, any binary T, the phylogenetic ideal I_T for the GM model is generated by edge invariants,

i.e., by all 3×3 minors of all matrix flattenings of P on edges of T.



Edge invariants for GM(k)

Focusing on edge e leads to a 'simpler' graphical model:



'Flattening' defines a map

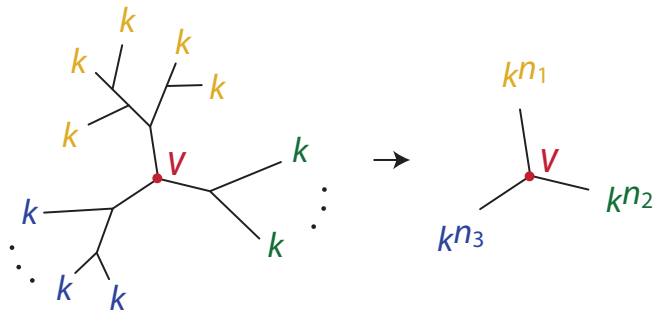
$$V_T \to \mathrm{Sec}^k(\mathbb{P}^{k^{n_1}-1} \times \mathbb{P}^{k^{n_2}-1})$$

i.e. to $k^{n_1} \times k^{n_2}$ matrices of rank at most k.

Edge invariants say that rank of the flattening is at most k.

Vertex invariants for GM(k)

For $k > 2$, there are also 'vertex invariants'



For $P \in V_T$, flatten to 3-dim tensor

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

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

(i.e. map $V_T \to \mathrm{Sec}^k(\mathbb{P}^{k^{n_1}-1} \times \mathbb{P}^{k^{n_2}-1} \times \mathbb{P}^{k^{n_3}-1})$;
vertex invariants belong to the ideal of the secant variety.)

From small trees to large trees

Main result for $k > 2$:

Theorem (A, Rhodes 2008):

For any k, given all invariants associated to the 3-taxon tree, we can explicitly construct set-theoretic defining polynomials for V_T for GM model on any binary tree T.

Extension to ideal generators. (Draisma, Kuttler 2010)

In the case of DNA, ($k = 4$), we still do not know generators of I_{T_3} for the 3-leaf tree.

Salmon problem

In light of this theorem, at the IMA in March 2007 proposed

Problem: Determine the ideal defining $\mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$

Reward: Smoked Copper River Salmon (caught just for you...)

2010, 2011 Friedland, Gross find(s) set-theoretic polynomials defining V_T. Later prove lower degree set suffices.

2010 Bates, Oeding using numerical methods (Bertini) find lower degree set.

II. Identifiability results

Parameters for a phylogenetic model are

identifiable

if the parameterization map ψ is $1-1$ (up to some understood symmetry).

That is, if $P \in \psi(\text{parameters})$, can we identify the parameters?

Identifiability asks

What parameters can be inferred under ideal circumstances?

and is necessary for sound statistical inference.

II. Identifiability results

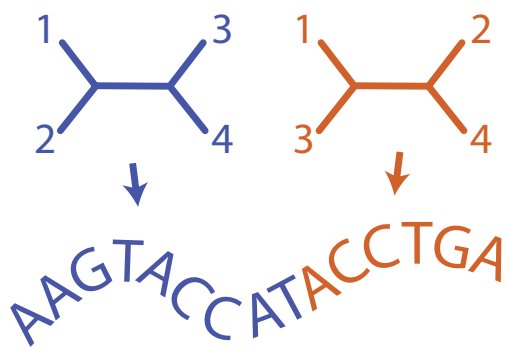
In a series of phylogenetics papers (2005-2008),

some negative (and scary) identifiability results were reported.

For example, under specialized substitution models

- Construction of a dist P arising from a mixture of two trees T_1, T_2 that *exactly matches* a distribution from a third tree T_3.

It is well accepted that genetic data can contain multiple tree signals (reticulate evolution, process dependent on position along genome, gene transfer, incomplete lineage sorting, etc.)



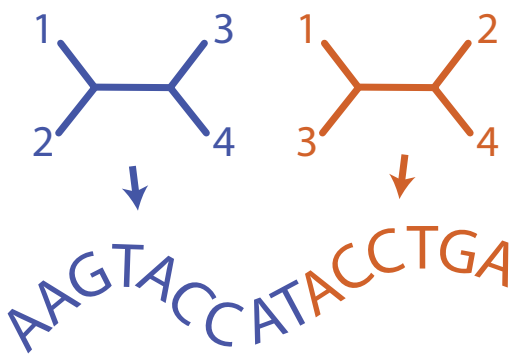
II. Identifiability results

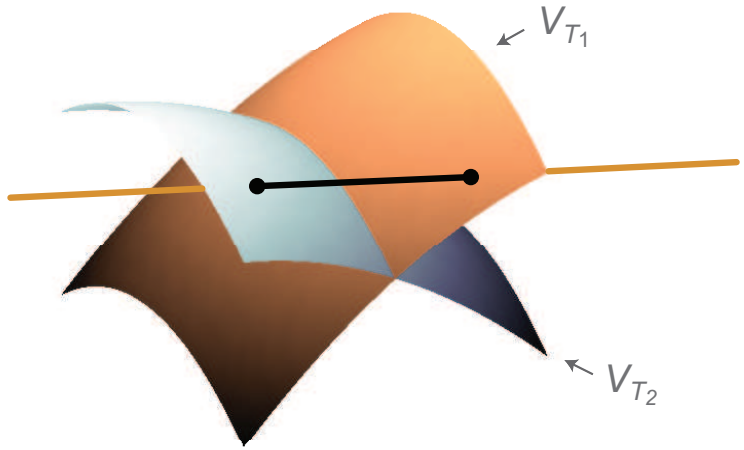
A 2-class mixture model can (begin to) model this.

Here a distribution P is the convex sum of dists for T_1 and T_2.

$$P = \delta P_{T_1} + (1-\delta) P_{T_2}$$

At variety level, this gives rise to a *join*, $\text{Join}(V_{T_1}, V_{T_2})$.





II. Identifiability results

If a distribution on two trees can exactly match a distribution from a single tree, then

it is **impossible** *to determine if data has*

single-tree signal (infer 'the' tree) or

multiple-tree signals (infer the trees or network).

Thus, it is impossible to determine the evolutionary relationships between taxa (in this statistical framework).

II. Identifiability results

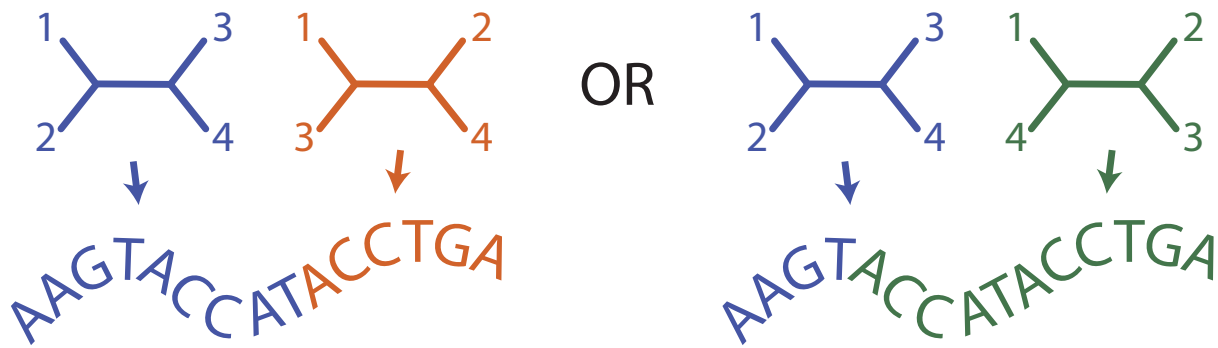
Another non-identifiability example:

- Proof that a mixture distribution P for trees T_1 and T_2 can *exactly match* a mixture distribution for trees T_3, T_4.

  Thus, since

$$P \in \text{Join}(V_{T_1}, V_{T_2}) \cap \text{Join}(V_{T_3}, V_{T_4}),$$

it is **impossible** *to infer trees in a mixture.*



II. Identifiability results

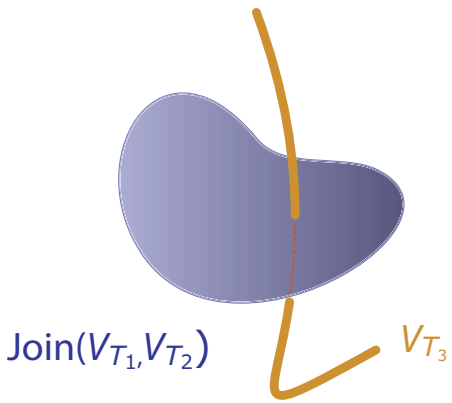
Better viewpoint:

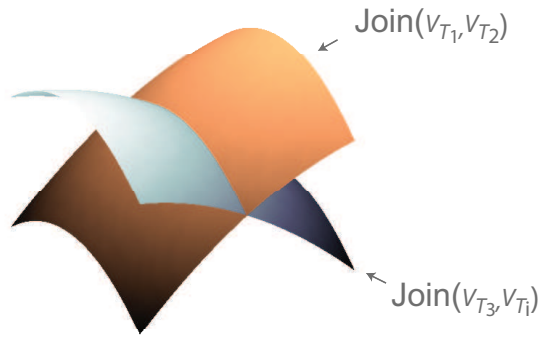
Consider the relevant varieties and investigate

generic identifiability.

Is a phylogenetic variety V_{T_3} contained in $\text{Join}(V_{T_1}, V_{T_2})$?

Are two irreducible phylogenetic varieties of the same dimension the same? or distinct?





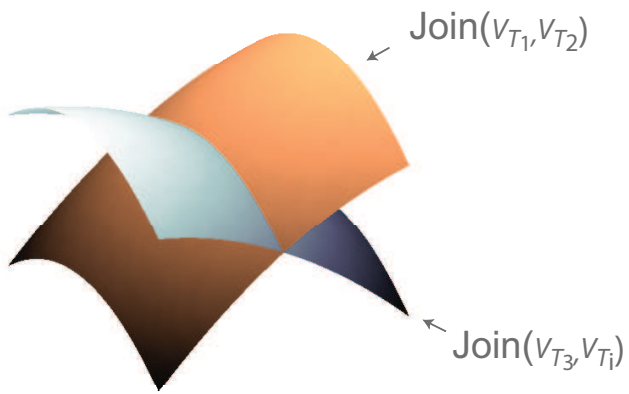
$$\exists f \in I_{\text{Join}(V_{T_1}, V_{T_2})} \setminus I_{V_{T_3}} \qquad I_{\text{Join}(V_{T_1}, V_{T_2})} \neq I_{\text{Join}(V_{T_3}, V_{T_i})}$$

II. Identifiability results

(A, Petrović, Rhodes, Sullivant 2011)

In recent work, using geometric methods we show that these scary examples were 'non-generic.'

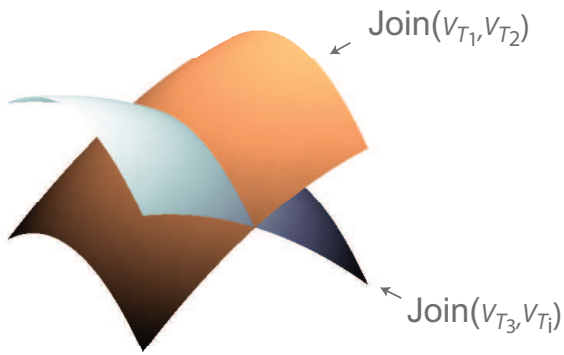
For a special class of models (group-based models) and a 2-class mixture model, both the tree parameters and numerical parameters are generically identifiable for n-taxon trees, if $n \geq 5$.



Method of identifiability proofs

Polynomial relationships between pattern probabilities reflect tree topology.

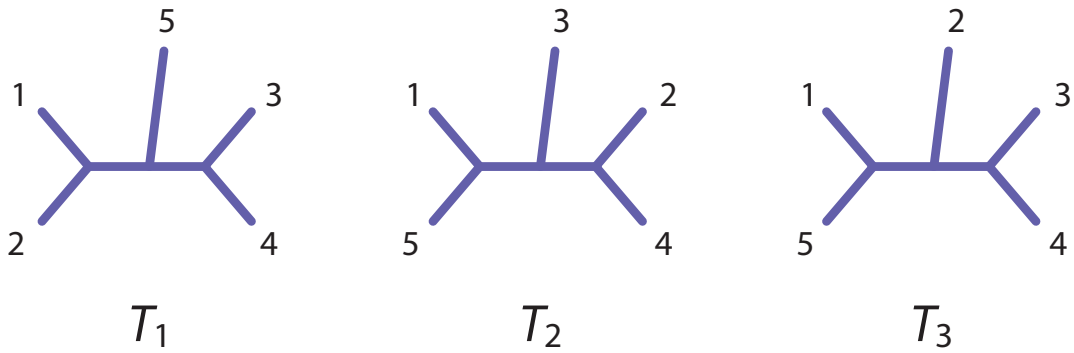
- Pattern probabilities for a fixed tree topology (or a mixture of two topologies) form a variety. More correctly, a semi-algebraic set.
- Varying the tree topologies changes the varieties.



If $\{T_1, T_2\} \neq \{T_3, T_4\}$, then the join varieties are distinct. This is proved by finding (polynomials in) their defining ideals.

- Understanding the variety and its ideal leads to insights for extracting information about the tree(s). Since the ideals are complicated, computational methods are helpful.

III. More identifiability results



Theorem (A, Petrović, Rhodes, Sullivant 2011)

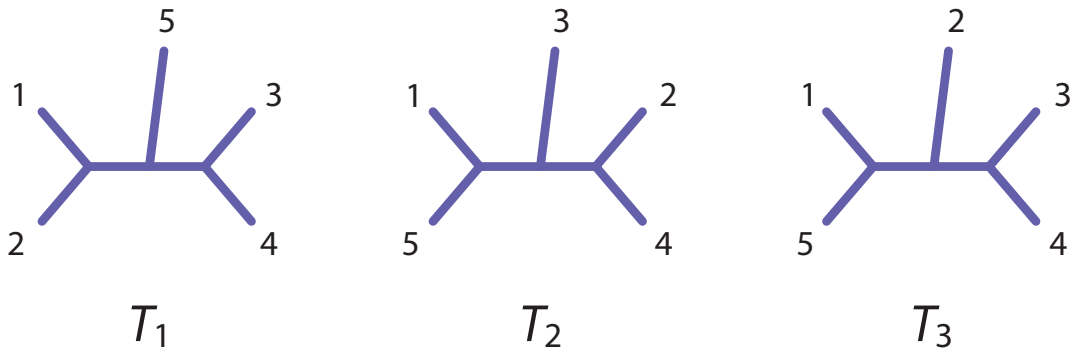
For the Jukes-Cantor model

$$V_{T_3} \subseteq \text{Join}(V_{T_1}, V_{T_2}),$$

where $\text{Join}(V_{T_1}, V_{T_2})$ corresponds to a mixture of 4-taxon trees.

Question: Is the same true for the image of stochastic parameters?

III. More identifiability results



Question: If $\mathcal{M}_i = \mathrm{Im}(\psi^{st}_{T_i})$

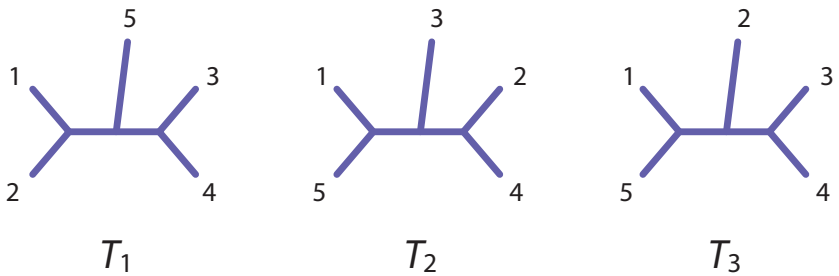
$$\text{Is } \mathcal{M}_3 \subseteq \mathrm{Join}^{st}(\mathcal{M}_1, \mathcal{M}_2)?$$

Note: This is a *semi-algebraic* question. (*real* algebraic geometry)

Can inequalities distinguish the stochastically parameterized dists, and do the same containments hold?

III. More identifiability results

Is $\mathcal{M}_3 \subseteq \text{Join}^{st}(\mathcal{M}_1, \mathcal{M}_2)$?

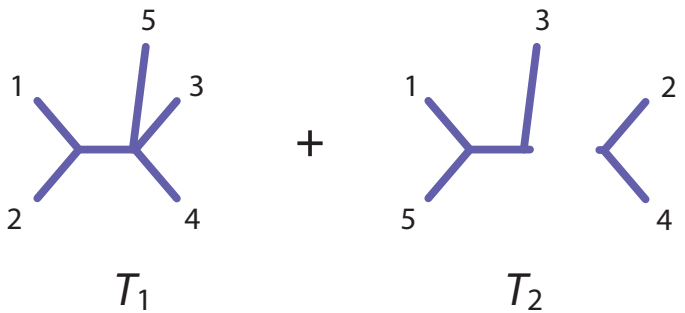


A*: No, unless allow 0 and/or infinite branch lengths in T_1 and T_2.

A* given by a *primary decomposition* computation of an ideal $\mathcal{I}$.

(Finds irreducible components of $V(\mathcal{I})$.)

One 2-dimensional component arises from mixtures like:



IV. Gene trees vs. species trees

```
Gorilla      AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCAT...
Orangutan    AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCT...
Human        AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCT...
Chimpanzee   AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT...
Gibbon       AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTT...
                                  ⋮
```

Given aligned DNA gene sequences from a number of different species, infer the evolutionary tree relating them.

them = the genes? the species?

These are not the same.

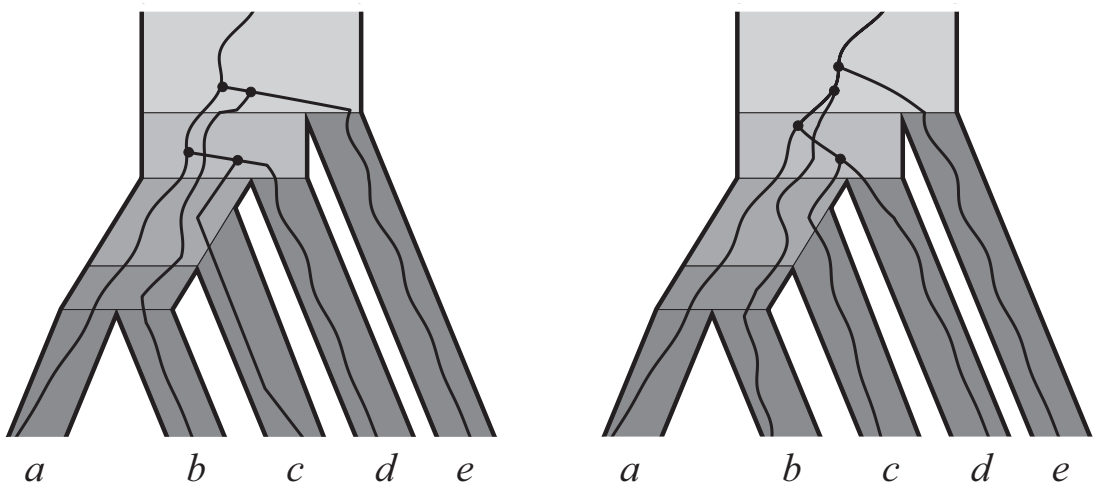
All that has been discussed in this talk concerns gene trees, but

gene trees may differ from the underlying species tree.

Gene tree discordance

Different gene trees for the same set of taxa often disagree.

- One cause is incomplete lineage sorting.



Gene trees $(((C, D), A), (B, E))$ and $((((C, D), A), B), E)$

in species tree $((((a, b), c), d), e)$

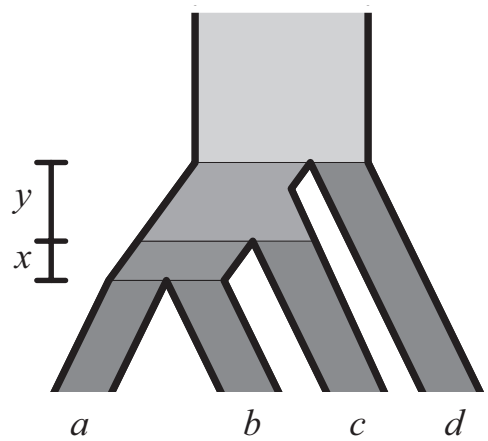
But, discordance of gene trees gives information about species tree.

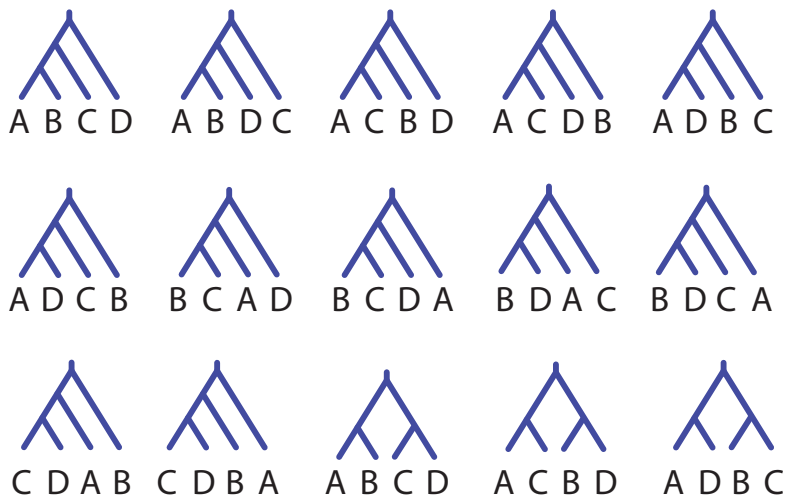
The Multispecies coalescent model

Incomplete lineage sorting is modeled by the

multispecies coalescent model.

This model gives the *gene tree distribution*, the theoretical proportions of rooted topological gene trees arising on a fixed, rooted, metric species tree.



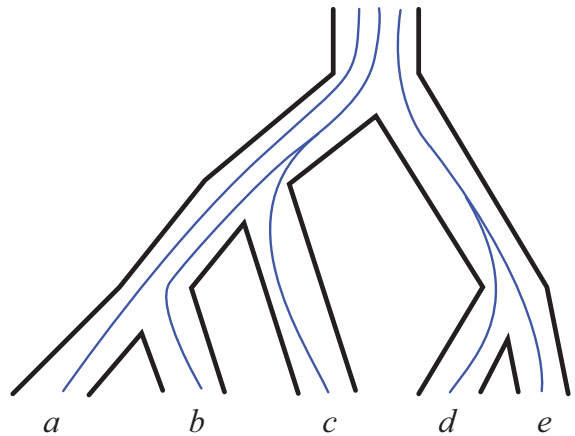


Notation: For a species tree σ, the leaves are labeled by $\mathcal{X} = \{a, b, \dots\}$. Gene trees use capital letters A, ...

The Multispecies coalescent model

- Viewing time backwards (present $\longrightarrow$ past), lineages within a population coalesce, one pair at a time.

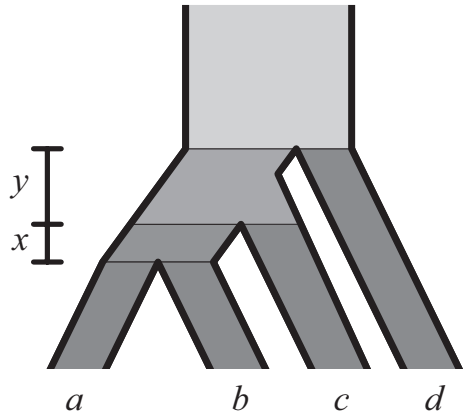
  The species tree constrains which lineages may coalesce at any given time.



- Given a species tree σ on $\mathcal{X}$, one can compute the (rooted, topological) *gene tree distribution*.

  This distribution is parameterized by the species tree topology and its internal branch lengths.

Gene tree distributions



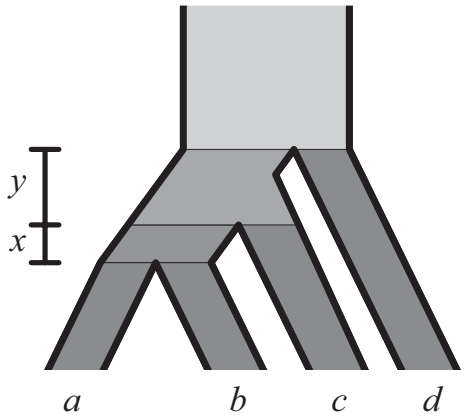
Coalescent events occur in 'populations,' or internal branches of species tree σ.

x, y in coalescent units ($\propto$ time/pop. size)

If k lineages enter a population from below, then coalescent events

- follow a Poisson process with rate $\binom{k}{2}$; Thus, the waiting time for a coalescent event in any population is modeled by an exponential random variable.
- any two lineages are equally likely to coalesce, with probability $\binom{k}{2}^{-1}$

Gene tree distributions



For this fixed σ, the multispecies coalescent yields the prob of the matching gene tree $P(G \mid \sigma)$ and the 14 other rooted gene tree probs.

$$P(G \mid \sigma) = 1 - \frac{2}{3}e^{-x} - \frac{2}{3}e^{-y} + \frac{1}{3}e^{-x}e^{-y} + \frac{1}{18}e^{-x}e^{-3y}$$

In general, for a n-taxon species tree σ, there are $(2n-3)!! = 1 \cdot 3 \cdot 5 \cdots (2n-3)$ rooted gene tree probabilities. Thus,

- Large and complicated distribution in $\exp(-x_i)$
- Using change of variable, $X_i = \exp(-x_i)$, this is a polynomial parameterization map.
- For a fixed species tree σ, this map defines a variety V_σ. V_σ is low dim variety in high dim ambient space.

Identifiability of species trees

Question: Given the set of rooted gene tree probabilities, can the species tree σ be identified?

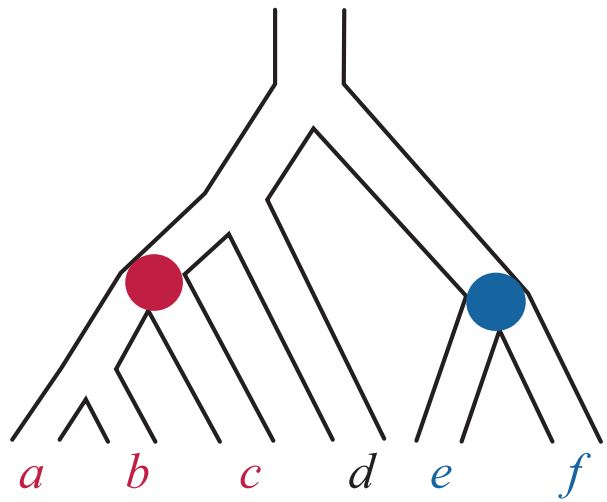
More generally, how can we summarize the gene tree dist and still retain enough information to recover the species tree σ?

The *species tree (including branch lengths) can be identified*

- from rooted triples probabilities, (Degnan, DeGiorgio, Bryant, Rosenberg 2009)
- from *unrooted* gene tree probabilities, by arguments based in algebra (A, Degnan, Rhodes, *to appear*)

Identifiability from clade probabilities

Defn: A *clade* $\mathcal{C}$ on a species tree σ is the collection of taxa all descended from a node of σ.



Clades $\mathcal{C}_1 = \{a, b, c\}$ and $\mathcal{C}_2 = \{e, f\}$ depicted at left.

A clade on a gene tree T is defined similarly.

Question:

Do clade probabilities on gene trees determine the species tree?

Clade probabilities

- summarize a collection of gene trees;

  Clade probability is proportion of gene trees displaying clade $\mathcal{C}$

- don't depend on metric information on gene trees

  Inferred branch lengths in metric gene trees can be sensitive to model choice (lack of robustness)

- *do* contain enough information to identify a species tree σ;

  We prove that if two species trees are different, $\sigma_1 \neq \sigma_2$, then the corresponding varieties are distinct.

Eg. Clade probabilities on 4-taxon trees

Probabilities of clades for 4-taxon species trees under the coalescent. $X = \exp(-x)$, $Y = \exp(-y)$.

clade	probability under species tree $(((a,b):x,c):y,d)$	$((a,b):x,(c,d):y)$
$c_1 = \Pr(AB)$	$1 - \frac{2}{3}X - \frac{1}{9}XY^3$	$1 - \frac{2}{3}X - \frac{1}{9}XY$
$c_2 = \Pr(AC)$	$\frac{1}{3}X - \frac{1}{9}XY^3$	$\frac{2}{9}XY$
$c_3 = \Pr(AD)$	$\frac{1}{6}XY + \frac{1}{18}XY^3$	$\frac{2}{9}XY$
$c_4 = \Pr(BC)$	$\frac{1}{3}X - \frac{1}{9}XY^3$	$\frac{2}{9}XY$
$c_5 = \Pr(BD)$	$\frac{1}{6}XY + \frac{1}{18}XY^3$	$\frac{2}{9}XY$
$c_6 = \Pr(CD)$	$\frac{1}{3}Y - \frac{1}{6}XY + \frac{1}{18}XY^3$	$1 - \frac{2}{3}Y - \frac{1}{9}XY$
$c_7 = \Pr(ABC)$	$1 - \frac{2}{3}Y - \frac{1}{3}XY + \frac{1}{6}XY^3$	$\frac{1}{3}Y - \frac{1}{6}XY$
$c_8 = \Pr(ABD)$	$\frac{1}{3}Y - \frac{1}{6}XY$	$\frac{1}{3}Y - \frac{1}{6}XY$
$c_9 = \Pr(ACD)$	$\frac{1}{6}XY$	$\frac{1}{3}X - \frac{1}{6}XY$
$c_{10} = \Pr(BCD)$	$\frac{1}{6}XY$	$\frac{1}{3}X - \frac{1}{6}XY$

These parameterizations give rise to varieties V_σ. (surfaces here)

Polynomials in I_{V_σ}

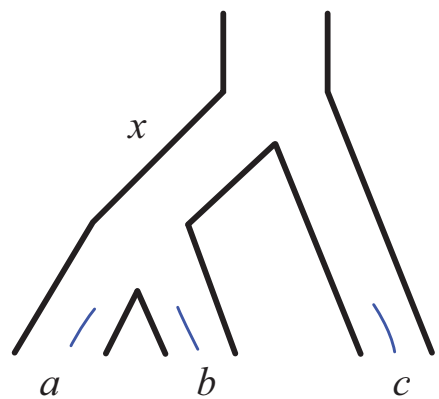
'Cherry swapping' invariants (eg: polynomials in ideal of V_σ)

Proposition. If (a, b) is a 2-clade on the species tree, then for any non-empty set $\mathcal{D}$ of taxa excluding A, B,

$$\Pr(\mathcal{D} \cup \{A\}) = \Pr(\mathcal{D} \cup \{B\}).$$

For example,

$\Pr(\{AC\}) - \Pr(\{BC\})$
is an invariant in I_{V_σ}.



An additional argument shows that we can identify 2-clades on σ.

Results (clade probs)

(A, Degnan, Rhodes, preprint)

With insight gained from extensive use of computational algebra methods, ...

we were able to generalize the 'cherry swapping' idea and to understand why certain polynomial relationships hold.

Theorem.

Clade probabilities determine the species tree topology.

For small trees, it is also possible to recover species tree branch lengths, but no general method is known.

Concluding remarks

- Statistical models in phylogenetics have a rich algebraic structure.
- The algebra/geometry viewpoint was crucial to gaining the insights and results described here, (understand limits of what can be inferred).
- This area is rich with possibilities for interactions between mathematicians and biologists (and statisticians and computer scientists and ....)

Collaborators

Joint work with:

John A. Rhodes
University of Alaska Fairbanks

Peter Jarvis
University of Tasmania

James Degnan
University of Canterbury, NZ

Amelia Taylor
Colorado College

Jeremy Sumner
University of Tasmania

Sonja Petrović
University of Illinois Chicago

Seth Sullivant
North Carolina State University