Phylogenetics and Algebraic Geometry: Problems from Biology

Elizabeth S. Allman
Department of Mathematics and Statistics
University of Southern Maine

John A. Rhodes
Department of Mathematics
Bates College

Porto Conte, Alghero, Sardegna, May 28, 2005

Problem:

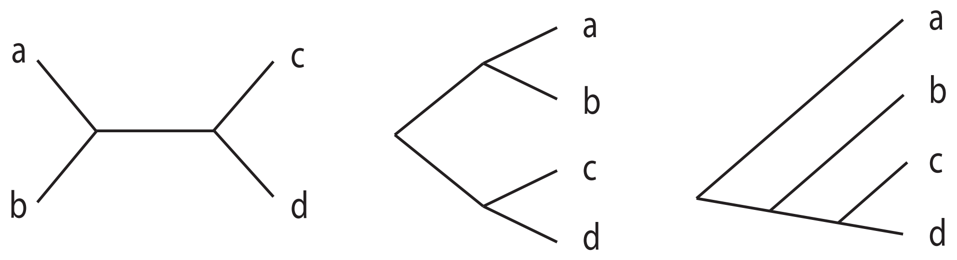
Given aligned biological sequences, presumed to have arisen from a common ancestral sequence, infer their evolutionary history.

a: AATCGCTGCTCGACC...

b: AAATGCTACTGGACC...

c: AAACGTTACTGGAGC...

d: AATCGTGGCTCGATC...



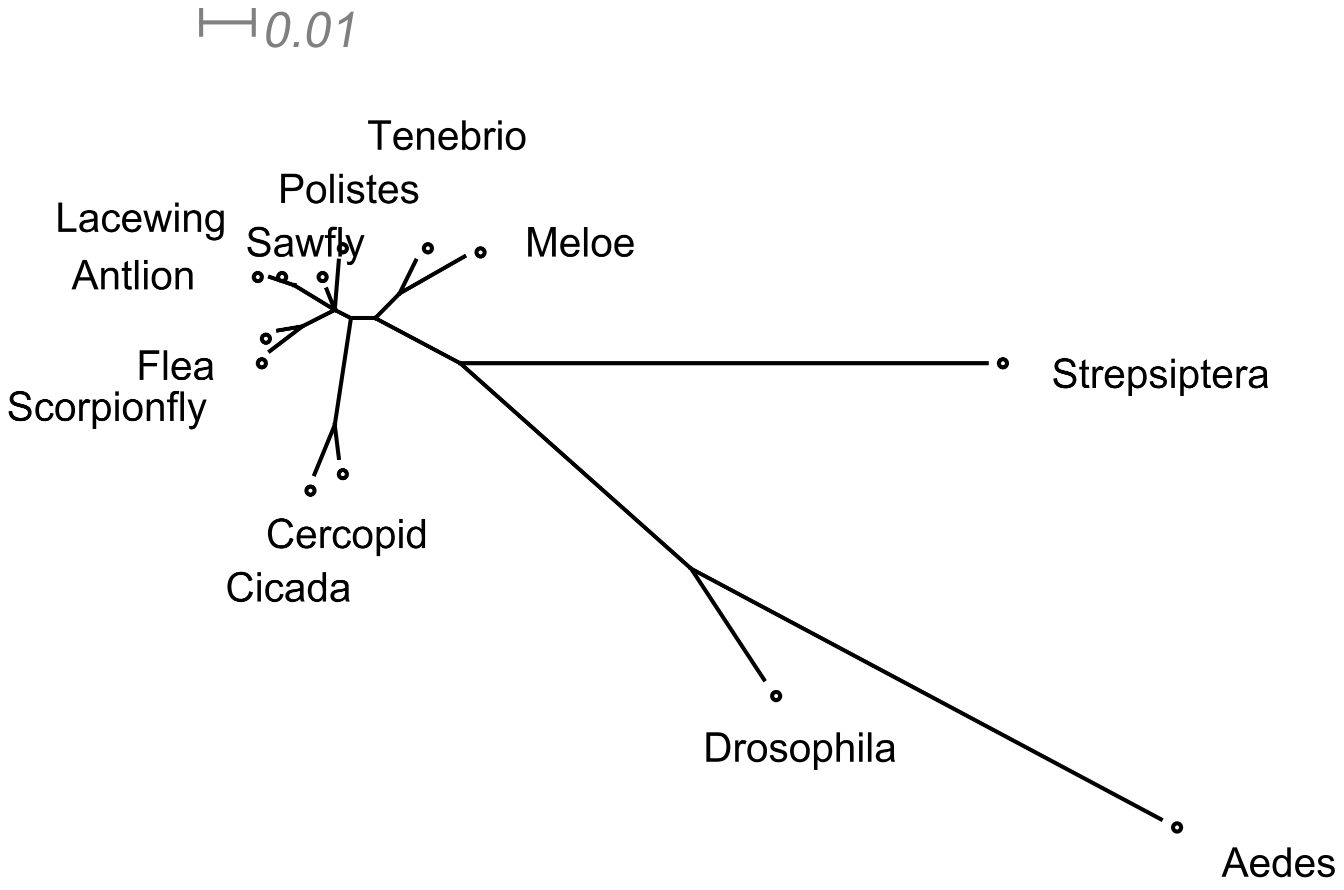
root location? sequences at internal nodes? edge lengths? description of mutation process along edges?

Example: 18S ribosomal DNA sequences, Insects

Strepsiptera	AAGCTCATTAAATCGCTTTGGTTCCTTAGATAGTTGGATAACTGTGGTAATTCTAGAGC...
Aedes	AGGCTCAGTATAACACTATAATTTACAAGATCATTGGATAACTGTGGAAAATCTAGAGC...
Drosophila	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Flea	TGGCTCATTATATCATTATGGTTCATTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Scorpionfly	TGGCTCATTACATCATTATGGTTCCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Lacewing	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Antlion	AGGCTCATTATATCATTATGGTTCCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Sawfly	TGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Meloe	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Polistes	TGGCTCATTAAATCATTATGGTTTCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Tenebrio	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...
Cicada	AGGCTCATTAAATCATTATGGTTCCTTGGATCTTTGGATAACTGTGGTAATTCTAGAGC...
Cercopid	AGGCTCATTAAATCATTATGGTTCCTTAGATCGTTGGATAACTGTGGTAATTCTAGAGC...

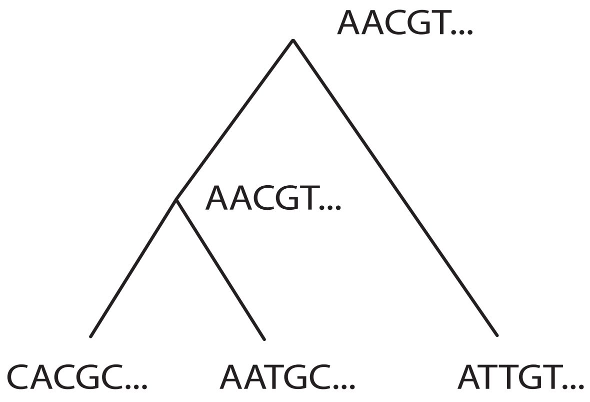
of length 770 sites, Neighbor Joining leads to....

Whiting, M.F., J.C. Carpenter, Q.D. Wheeler, and W.C. Wheeler. *Syst. Biol.* (1997) 46:1-68.



NJ, log-det distance (SplitsTree4, Huson and Bryant, 2004)

Probabilistic model of molecular evolution:

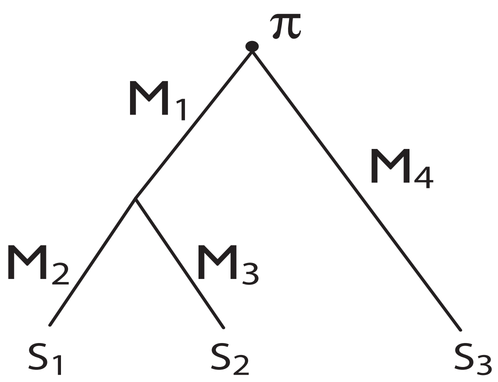


Description of process at a single site:

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

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

This is the general Markov model — GM — on the tree T.



Given T, $\boldsymbol{\pi}$,$\{M_e\}$, compute joint distribution of bases at leaves:
E.g., $GAA \rightsquigarrow 311$,

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

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

each p_{ijk} is polynomial in unknown parameters.

For T with n leaves, sequences with κ bases

- the joint distribution P is an n-dimensional $\kappa \times \kappa \times \cdots \times \kappa$ tensor.
- entries of P are polynomials in entries of π, $\{M_e\}$
- these polynomials reflect the topology of T
- for trivalent tree there are $N = (\kappa - 1) + (2n - 3)\kappa(\kappa - 1)$ parameters

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

A biological inference problem:

From aligned sequence data, first estimate joint distribution tensor $P = (p_{ijk\dots})$ by counting occurrences of base patterns.

a: ATTAGGTACATGATTAG

b: ATTCGGTACATGATTAG

c: ATTCGCTACATGATCCG

d: ATTTGCTACATGTTCCG

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

Then use the estimate $\widehat{P}$ to infer the topology of the evolutionary tree T, assuming a model such as GM.

Note that none of T, π, $\{M_e\}$ are known; but biologists care most about T.

A mathematical problem:

Since ϕ_T is polynomial, extend to a polynomial map

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

Use algebraic geometry to understand the image, the *phylogenetic variety*,

$$V_T = \overline{\phi_T(\mathbb{C}^N)}.$$

Since $\kappa^n >> N$, the pattern frequencies p_{ijkl} will satisfy polynomial relations. These equations are called *phylogenetic invariants* or model invariants for (T, GM).

Finding invariants $\leftrightsquigarrow$ finding an implicit description of V_T,

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

i.e. finding the kernel of

$$\Phi_T : \mathbb{C}[p_{0\ldots 0}, \cdots, p_{\kappa\ldots\kappa}] \longrightarrow \mathbb{C}[s_1, \cdots, s_N]$$

$$\ker \Phi_T = I_T \equiv \textit{ phylogenetic ideal,}$$

the ideal of polynomials in $p_{0\ldots 0}, \ldots, p_{\kappa\ldots\kappa}$ vanishing for all choices of (complex) parameters.

Only one invariant is easy to see – stochastic invariant

$$1 - \sum_{ijkl} p_{ijkl}$$

For small trees other invariants can be determined computationally.

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

Ex: GM model, $\kappa = 4$, for 3 or more leaves, lowest degree invariants are of degree 5, 180 summands....

In 1987,

Cavender and Felsenstein (JC)

Lake (‘K2P’)

proposed using invariants for phylogenetic inference.

Idea is to evaluate invariants at 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, GM, are the correct tree and mutation model relating the sequences, then $\widehat{P} \approx P = \phi(s) \in V_T$, for some parameters s.

For $f \in I_T$, $f(P) = 0$, so $f(\widehat{P}) \approx 0$

Implementation:

- Find invariants
- return tree for which $\widehat{P}$ "$\in$" V_T (as best possible)

Method is statistically consistent.

More generators of I_T in hand $\longleftrightarrow$ improved tree inference.

Issues:

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)

Basic Problem:

For any fixed tree T and κ, find all invariants.

- Ad hoc methods
- Gröbner basis techniques on small trees, simple models
- Recent work: to be described...

There are many variations on the model —

Number of bases:

- $\kappa = 4$, DNA A, T, C, G
- $\kappa = 2$, purine/pyrimidine $R = \{A, G\}, Y = \{C, T\}$
- $\kappa = 20$, proteins are sequences built from 20 amino acids

Special forms for $\boldsymbol{\pi}, \{M_e\}$

- Jukes-Cantor (1-parameter per edge)

$$\boldsymbol{\pi} = (.25 \quad .25 \quad .25 \quad .25),$$

$$M_e = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix}$$

- Kimura (group-based) model (3 parameters per edge)

$$\boldsymbol{\pi} = (.25 \quad .25 \quad .25 \quad .25),$$

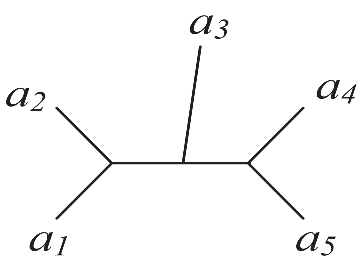
$$M_e = \begin{pmatrix} 1-\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & 1-\alpha-\beta-\gamma & \gamma & \beta \\ \beta & \gamma & 1-\alpha-\beta-\gamma & \alpha \\ \gamma & \beta & \alpha & 1-\alpha-\beta-\gamma \end{pmatrix}$$

For these models, work of Hendy, Hendy and Penny, Steel-Széleky-Erdös, Evans-Speed recognized role of Fourier transform (Hadamard conjugation).

Then Sturmfels-Sullivant recognized this means that the variety V_T is toric, and completed determination of the ideal I_T.

For GM model,

- $\kappa = 2$ ideal is known (AR),
- $\kappa > 2$ is partially understood (AR).



Example: $\kappa = 2$, GM

P, a $2 \times 2 \times 2 \times 2 \times 2$ array,

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

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

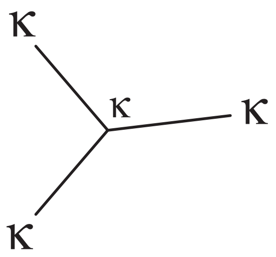
and

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

Theorem (Conjecture of Pachter-Sturmfels): For $\kappa = 2$ the ideal $I_T = I(V_T)$ of phylogenetic invariants for GM model on this T is generated by all 3×3 minors of these two matrices, and similarly for other trivalent trees.

Ideas behind this and related theorems...

For any κ, if T has 3 leaves



$V_T = V(\kappa; \kappa, \kappa, \kappa)$

$$p_{ijk} = \sum_l \pi_l M_1(l,i) M_2(l,j) M_3(l,k)$$

But $M_e(l, \cdot) \in \mathbb{P}^{\kappa-1}$, so

$$\begin{aligned} V(\kappa; \kappa, \kappa, \kappa) &= \mathrm{Sec}^{\kappa}(\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1}) \\ &= \kappa \times \kappa \times \kappa \text{ tensors of rank} \leq \kappa \end{aligned}$$

This makes the problem classical — but doesn't solve it.

Known:

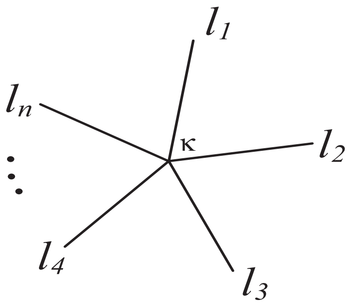
- $V(2;2,2,2) = \mathbb{P}^7$, defining ideal is (0)

- $V(3;3,3,3)$, ideal is generated by 27 quartics, constructed by Strassen, AR03; shown to generate by Garcia-Stillman-Sturmfels (computationally)

- $V(4;4,4,4)$, ideal requires 1728 quintics (Hagedorn, Landsburg-Manivel), constructed in AR03; also some degree nine generators are needed, constructed by Strassen; others?

- Many $\kappa + 1$ degree invariants for $V(\kappa;\kappa,\kappa,\kappa)$ were constructed by AR03.

Similar model on star trees with more leaves are also of interest for other statistical models.



$$V(\kappa; l_1, \ldots, l_n) = \mathrm{Sec}^{\kappa}(\mathbb{P}^{l_1-1} \times \cdots \times \mathbb{P}^{l_n-1})$$

Conjecture (Garcia-Stillman-Sturmfels): The full ideal defining

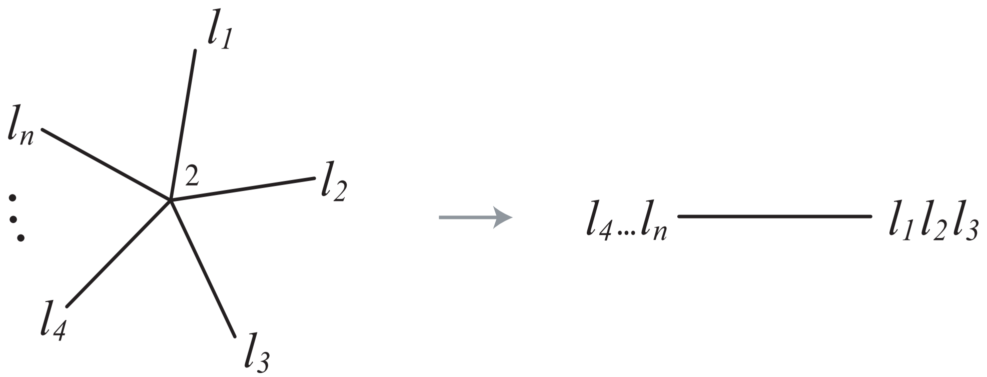
$$V(2; l_1, \ldots, l_n) = \mathrm{Sec}(\mathbb{P}^{l_1 - 1} \times \cdots \times \mathbb{P}^{l_n - 1})$$

is the sum of the ideals defining

$$V(2; l_1 l_2 \cdots l_k,\, l_{k+1} \cdots l_n) = \mathrm{Sec}(\mathbb{P}^{***} \times \mathbb{P}^{***})$$

(with permutations of the l_i).

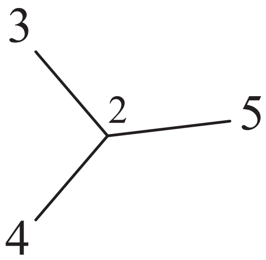
I.e., the ideal is generated by 3×3 minors of 2-d flattenings of a $l_1 \times l_2 \times \cdots \times l_n$ tensor.



Example: $V(2;3,4,5)$

A $3 \times 4 \times 5$ tensor flattens 3 ways: to 3×20, 4×15, and 5×12 matrices.

All 3×3 minors of these three matrices generate the ideal.



Previous results on GSS conjecture:

- GSS checked small cases computationally, $n \leq 5$.
- Landsberg-Manivel (via Weyman): $n = 3$ case

Theorem (AR): If GSS holds for $V(2; 2, 2, \ldots, 2)$, it holds for $V(2; l_1, l_2, \ldots, l_n)$.

Corollary: The GSS conjecture holds for $n \leq 5$.

Thus explicit generators can be given for the ideal vanishing on the secant variety of the Segre product of up to 5 projective spaces.

This is a special case of

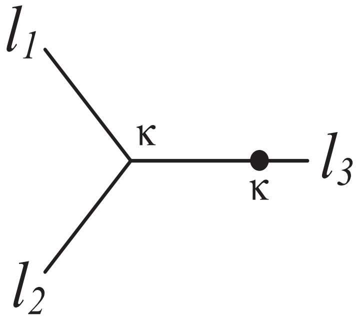
Theorem (AR): If $l_1, l_2, \dots, l_n \geq \kappa$, then generators of the ideal defining $V(\kappa; l_1, l_2, \dots, l_n)$ can be explicitly constructed from generators of the ideal defining $V(\kappa; \kappa, \kappa, \dots, \kappa)$.

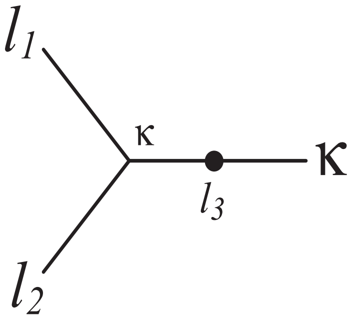
A glimpse of the proof: Observe that

$$V(\kappa; l_1, l_2, \kappa) *_{3,1} M_{\kappa \times l_3} = V(\kappa; l_1, l_2, l_3),$$

$$V(\kappa; l_1, l_2, l_3) *_{3,1} M_{l_3 \times \kappa} = V(\kappa; l_1, l_2, \kappa).$$

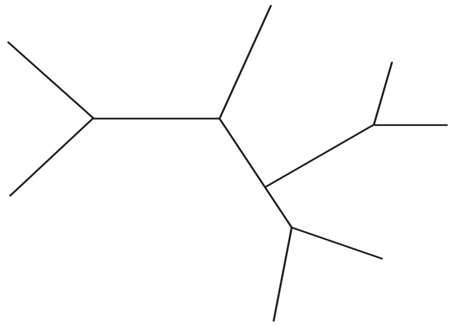
Here $M_{m\times n}$ denotes $m \times n$ matrices, and $*_{3,1}$ denotes 'matrix multiplication' in the 3 and 1 indices.





Get maps between ideals, related to $GL(l_3)$-action, and careful use of basic representation theory gives result.

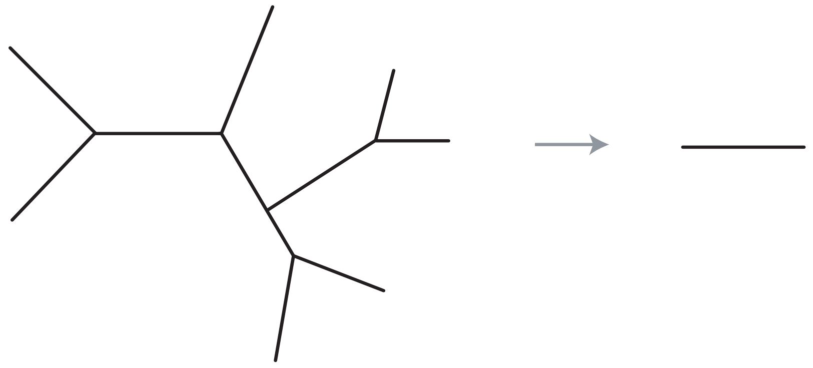
Back to general trees: n-leaf T



V_T depends on model and *topology* of T — but we can hope the ideal (or set-theoretic defining polynomials) can be described in terms of *local structure* of T.

How can local structure give ideal?

Place root on edge of T. Then 'collapse' tree, to a 'coarser' model $V(\kappa; \kappa^m, \kappa^{n-m})$.

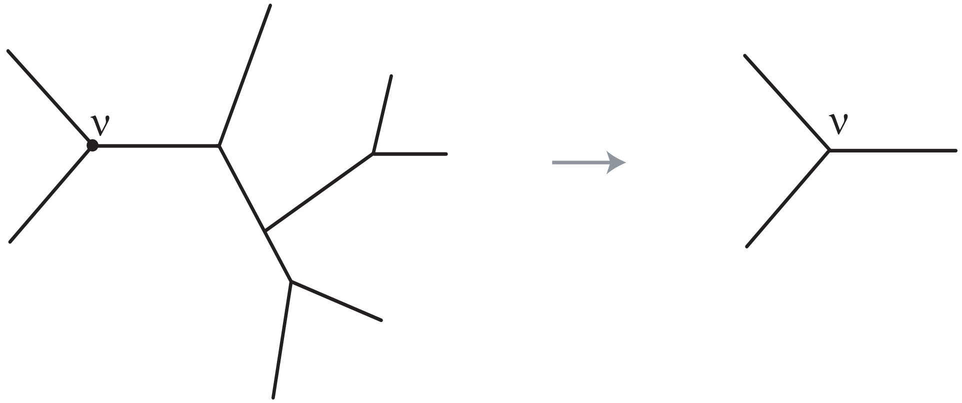


But

$$V(\kappa; \kappa^m, \kappa^{n-m}) = \mathrm{Sec}^\kappa(\mathbb{P}^{\kappa^m - 1} \times \mathbb{P}^{\kappa^{n-m}-1})$$
$$= \ \kappa^m \times \kappa^{n-m} \text{ matrices of rank } \leq \kappa.$$

Thus ideal generators are known for this model: $(\kappa + 1) \times (\kappa + 1)$ minors. These polynomials, the *edge invariants*, vanish on V_T.

Similarly there are *vertex invariants*:



Coarse model at v gives $V(\kappa; \kappa^{n_1}, \kappa^{n_2}, \kappa^{n_3})$, again a secant variety of a product of projective spaces.

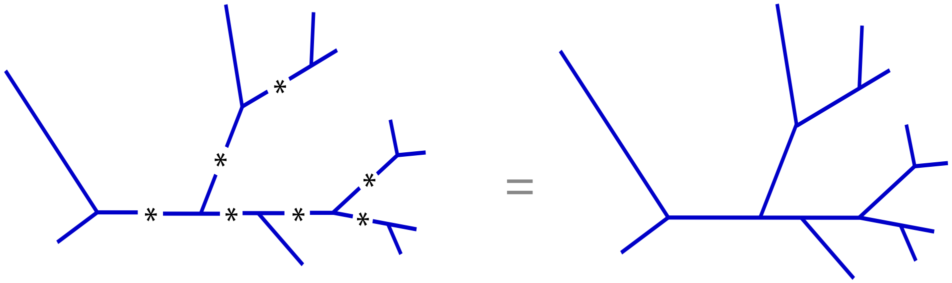
Two main results:

Theorem: For any κ, given *set-theoretic* defining polynomials of $V(\kappa;\kappa,\kappa,\kappa)$, we can explicitly construct *set-theoretic* defining polynomials for V_T for GM model on any trivalent tree T.

Main element of proof is

edge-invariants = matrix rank condition,

so can decompose tensor in V_T into product of tensors from smaller trees.



Theorem: For $\kappa = 2$, the ideal defining V_T *is generated* by edge invariants, i.e., by 3×3 minors of edge flattenings of the n-dimensional joint distribution tensor.

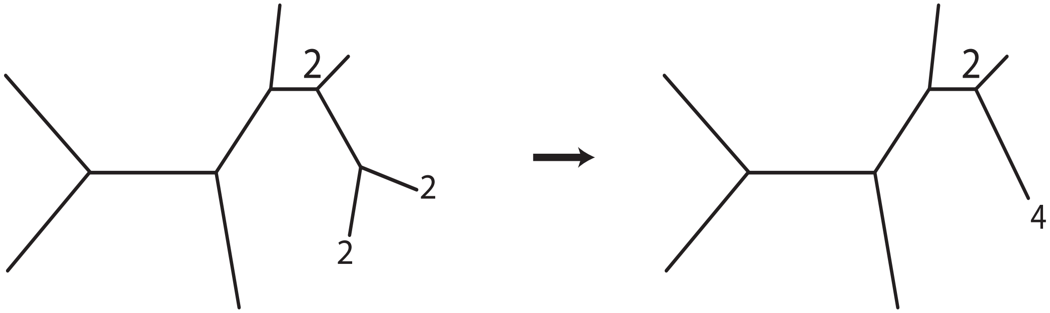
The proof uses the fact that $V(2; 2, 2, 2) = \mathbb{P}^7$ in two ways:

First, ideal for $V(2; 2, 2, 2)$ is (0), so generators for $V(2; 2^{n_1}, 2^{n_2}, 2^{n_3})$ are just edge invariants.

More importantly, $V(2; 2, 2, 2) = M_{2\times 4}$, so

$$V_T = V_{T'} * V(2; 2, 2, 2) = V_{T'} * M_{2\times 4}$$

so



Outstanding questions for GM:

- Determine ideal (or even set-theoretic defining polynomials) for $V(4;4,4,4) = \mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$, or more generally for $V(\kappa;\kappa,\kappa,\kappa)$ when $\kappa \geq 4$.

- Determine ideal for V_T for $\kappa = 4$, or more generally $\kappa \geq 3$, for arbitrary T.

- Determine ideal (or even set-theoretic defining polynomials) for

$$V(2;2,2,2,\ldots,2) = \mathrm{Sec}(\mathbb{P}^1 \times \cdots \times \mathbb{P}^1)$$

for 6 or more leaves/$\mathbb{P}^1$s.

Other models also need analysis ... and results short of determining ideal can be valuable.

- Covarion model

  8 states at internal nodes

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

  only 4 observable states at leaves

$$A, C, G, T$$

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

This model is believed to be more biologically realistic.

Algebraic viewpoint leads to:

Theorem: The topology of T is identifiable from a generic joint distribution tensor arising from the covarion model, using only 4-taxon comparisons.

(This is important for showing the Maximum Likelihood statistical method is consistent for the covarion model.)

- GM+I model

  2 classes of sites, one mutates according to GM, other is Invariable

  unknown which sites are in which class, unknown sizes of classes

Theorem: The topology of T is identifiable for GM+I using 4-taxon (and no fewer) comparisons. An explicit rational formula gives fraction of invariable sites.

- and many other models

  Stable base distribution (AR)

  Symmetric Strand model – interesting amalgam of group-based/GM models (Casanellas-Sullivant)

References:

Current Work:

Phylogenetic ideals and varieties for the general Markov model. 2004. preprint, `arXiv:math.AG/0410604`.

Phylogenetic invariants and parameter recovery for the general Markov plus invariable sites model. 2005. in preparation.

The identifiability of a general phylogenetic model, with application to the covarion model. 2005. in preparation.

Introductions to using algebraic geometry in phylogenetics and graphical models:

Nicholas Eriksson, Kristian Ranestad, Bernd Sturmfels, and Seth Sullivant. Phylogenetic Algebraic Geometry. *Proceedings of Projective Varieties with Unexpected Properties*, 2004. to appear. `arXiv:math.AG/0407033`.

Luis David Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of Bayesian networks. *J. Symbolic Comp.*, to appear. `arXiv:math.AG/0301255`

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

Background Papers on Invariants and Phylogenetic Inference:

Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.

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

Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, 2004.

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

Other Papers on Invariants and Phylogenetic Inference using Geometric Approach:

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

Marta Casanellas and Seth Sullivant. *The strand symmetric model*. in Algebraic Statistics for Computational Biology, ed. L. Pachter and B. Sturmfels, Cambridge University Press, 2005. to appear.

Nicholas Eriksson. Tree Construction Using Singular Value Decomposition. in Algebraic Statistics for Computational Biology, ed. L. Pachter and B. Sturmfels, Cambridge University Press, 2005. to appear.

Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 2004. to appear. `arXiv:q-bio.PE/0402015`.

eallman@maine.edu

`http://www.usm.maine.edu/~math/P-Allman.html`

jrhodes@bates.edu

`http://www.bates.edu/~jrhodes`