Phylogenetic Models: Algebra and Evolution

Elizabeth S. Allman

Dept. of Mathematics and Statistics
University of Alaska Fairbanks

Applications in biology, dynamics, and statistics

Institute for Mathematics and its Applications – Minneapolis, MN
March 6, 2007

Outline:

1. the inference problem

   DNA sequences $\rightsquigarrow$ evolutionary tree

2. sequence evolution

   probabilistic models on trees

3. phylogenetic ideals and varieties

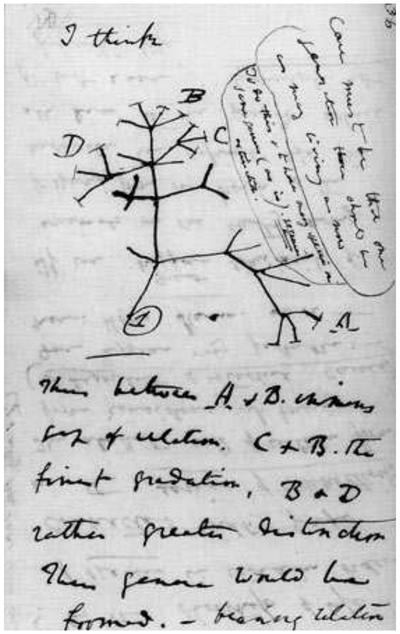
   models $\longleftrightarrow$ algebraic varieties

4. application

   identifiability of models

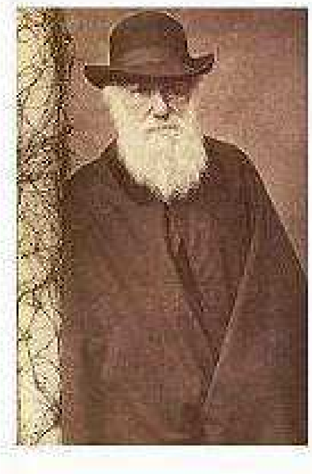
Inference Problem:

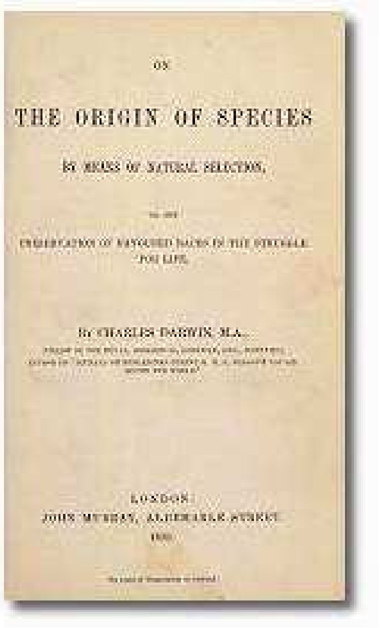
Given aligned biological sequences, presumed to have arisen from a common ancestral sequence, infer their evolutionary history.


I think
B
C
D
1
A

Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.

Thus between A & B immense gap of relation. C & B the finest gradation, B & D rather greater distinction. Thus genera would be formed. — bearing relation





ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

BY CHARLES DARWIN, M.A.,

[illegible]

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

[illegible]

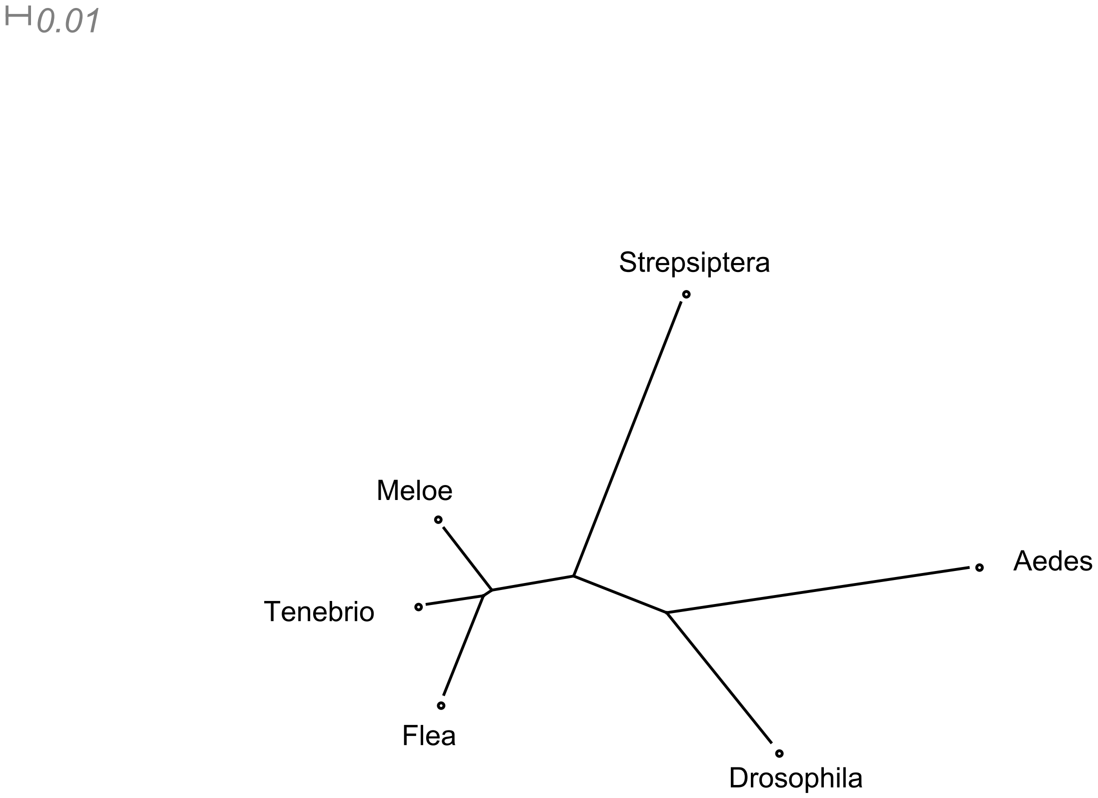
For phylogenetic inference,

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 true joint distribution p_{AAAAAA}, etc.


0.01
Strepsiptera
Meloe
Aedes
Tenebrio
Flea
Drosophila

Model-based methods.

With a probabilistic model of the mutation process specified, use

Statistical Frameworks:

- Maximum Likelihood – find the parameters Θ (especially the tree) that maximize the likelihood function, $L(\Theta) = P(\textit{data} \mid \Theta)$
- Bayesian Methods – find the posterior distribution on the parameters Θ (especially the tree)

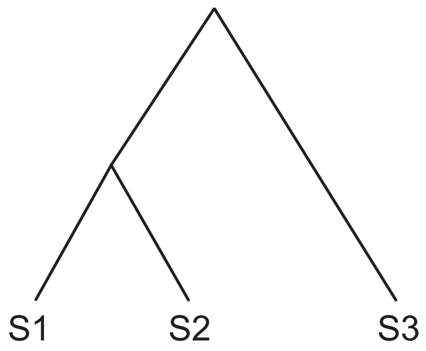
Software: PAUP*, Phylip, PAML, SplitsTree, Mr. Bayes, etc.

Modeling molecular evolution along a tree T:

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

root $=$ most recent common ancestor

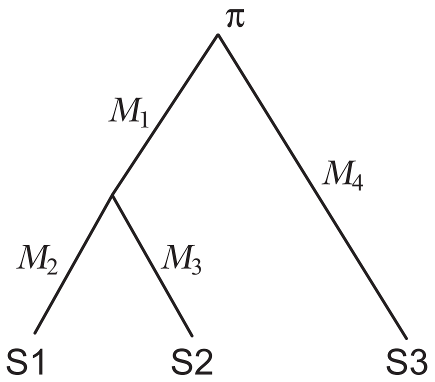
leaves $=$ currently extant taxa



κ states at each node,

$\kappa = 4$ (A,C,G,T), $\kappa = 20$ (proteins)

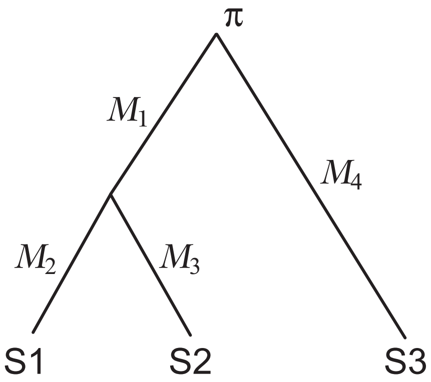
$\kappa = 2$ (R={A,G},Y={C,T}), $\kappa = 61$ (codons=triplets of A,C,G,T)



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

Model describes evolution at a single site in sequence

More specifically,



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

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

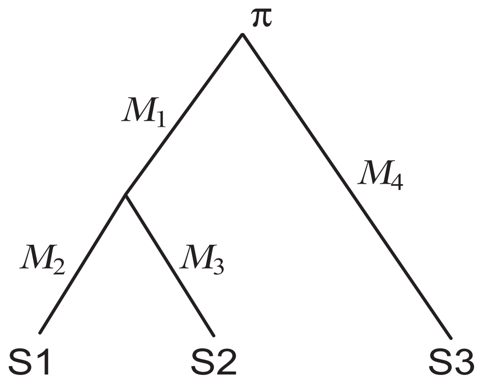
This is the general Markov model (GM) on the tree T.

(Other models $\mathcal{M}$ will appear later....)

Model parameters T, $\boldsymbol{\pi}$,$\{M_e\}$ lead to values for the

joint distribution of states at leaves

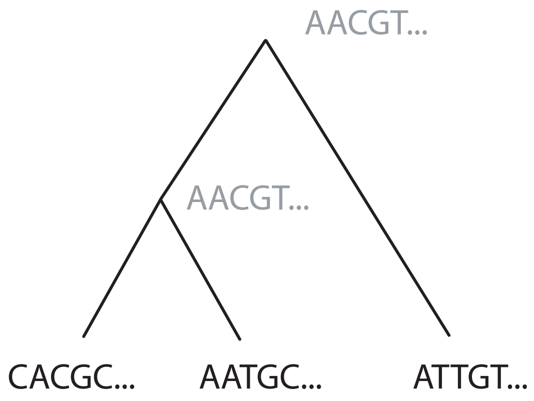
($=$ expected pattern frequencies).



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

$P = (p_{ijk})$ is a $\kappa \times \kappa \times \kappa$ tensor (table) with entries that

- are polynomial in the stochastic parameters
- can be estimated from data by $\widehat{p}_{ijk}$.



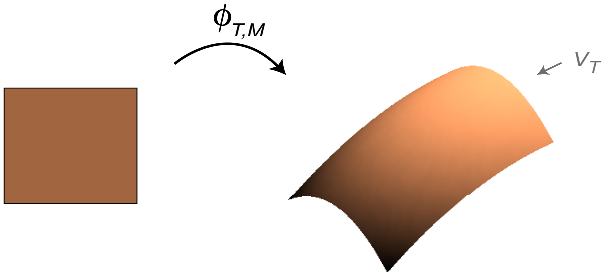
Note:

- Multiple sites, assume i.i.d.
- Data comes only from living taxa at leaves; states at internal nodes are *hidden*. (latent variables)
- Given state at any node, processes on descending edges are independent. (conditional independence)

For a fixed tree T, we have the polynomial map $\phi_T : (\boldsymbol{\pi}, \{M_e\}) \mapsto P$, which can be extended to the complex setting,

$$\phi_T : \{\text{Parameters on } T\} \longrightarrow \mathbb{C}^{\kappa^n}$$

The closure of the image, $\overline{\mathrm{Im}(\phi_T)}$, is the *phylogenetic variety V_T*.



This associates to each tree T an algebraic variety V_T whose points 'are' all joint distributions describing sequences that evolved along T.

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

I_T = the phylogenetic ideal

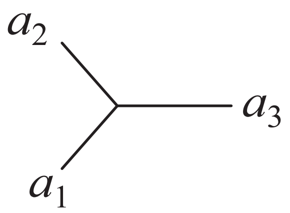
$f \in I_T$ is called a phylogenetic invariant

$$f \in I_T \Longleftrightarrow f(P) = 0 \text{ for all } P = \phi_T(\pi, \{M_e\})$$

Invariants depend on the topology of T and the choice of substitution model $\mathcal{M}$.

In principle, invariants can be computed — Gröbner bases, elimination; in practice, usually not.

Example: 3 taxa, GM, $\kappa = 4$ (“Bernd's favorite statistical model”)



Trivial invariant: $\sum p_{ijk} - 1$

There are no homogeneous invariants of degree < 5.

A 1728-dim space of *all* quintics in I_T can be explicitly constructed.

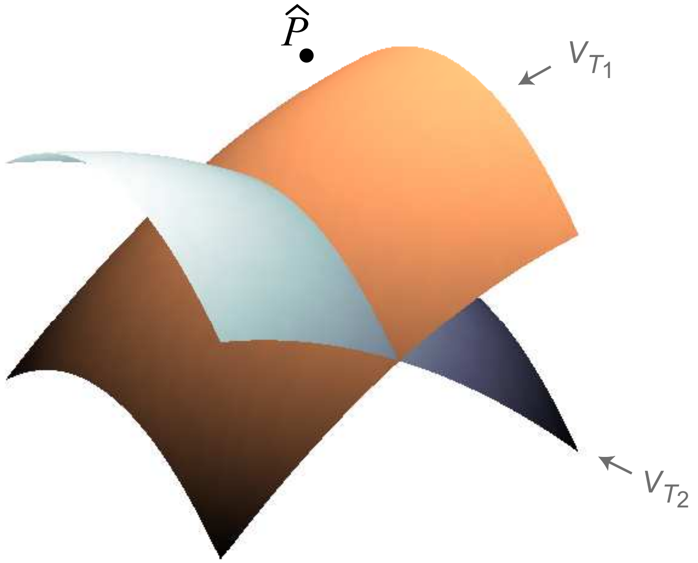
For instance ...

$$
\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} \\
& +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}
$$

$$
\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}
$$

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

Invariants were originally introduced for inference....



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

Models of sequence evolution

- group-based models: $\begin{cases}\text{Jukes-Cantor}\\ \text{Kimura models, K2P, K3ST}\end{cases}$
- general Markov models: GM
- continuous-time models: GTR – general time reversible
- Elaborations:

    mixture models: GM+I, GM+GM+GM, GTR+I,...

    rates-across-sites models: covarion, GTR+I+Γ

    2-tree mixtures,...

Group-based models (Jukes-Cantor, K2P, K3P):

Kimura 2-parameter model (2 parameters per edge – a, b)

$$\boldsymbol{\pi} = (.25 \quad .25 \quad .25 \quad .25), \quad M_{K2P} = \begin{pmatrix} * & a & b & b \\ a & * & b & b \\ b & b & * & a \\ b & b & a & * \end{pmatrix}$$

Rich algebraic structure (Hendy, Evans-Speed, Sturmfels-Sullivant,...)

After a change of coordinates (Fourier/Hadamard), parameterization map ϕ_T is given by monomials, i.e., V_T is toric.

I_T is well-understood (invariants tied to local features in trees: edges and nodes)

General Markov model (GM):

$$\text{arbitrary } \boldsymbol{\pi},\ \{M_e\}$$

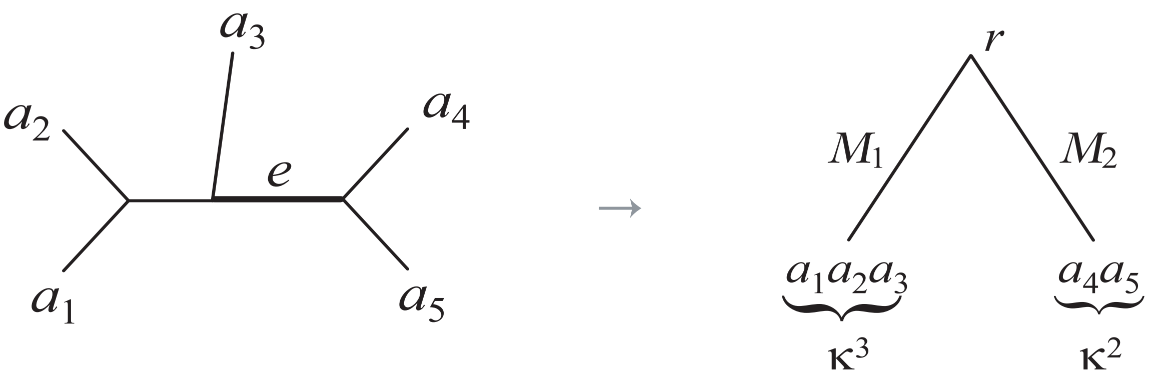
V_T is *not* toric.

Understanding of the I_T comes from rank conditions on matrices and tensors.

To the extent of current understanding, again invariants arise from local structure in trees.

Local nature of edge invariants...

Focusing on edge e leads to a ‘simpler’ graphical model:


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

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 $\mathrm{Flat}_e(P)$

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

$$\begin{aligned}\mathrm{Flat}_e(P) = M_1^T\,\mathrm{diag}(\boldsymbol{\pi}_r)M_2 &\\ &\iff \mathrm{rank}(\mathrm{Flat}_e(P)) \leq \kappa \\ &\iff \text{all } (\kappa+1)\times(\kappa+1) \text{ minors vanish.}\end{aligned}$$

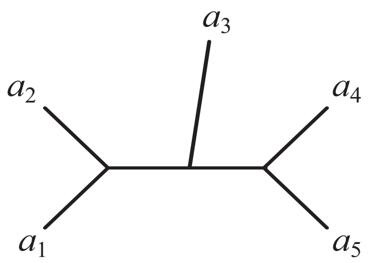
These minors are the edge invariants for a tree T.

$$\Big(\text{ i.e., } \mathrm{Flat}_e(P) \in \mathrm{Sec}^{\kappa}(\mathbb{P}^{\kappa^{n_1}-1} \times \mathbb{P}^{\kappa^{n_2}-1})\ \Big)$$

“The art of giving a different name to the same thing.”

$$\mathrm{Sec}^{\kappa} \leftrightsquigarrow \mathrm{Sec}^{\kappa-1}$$

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:

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

The corresponding flattenings are

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

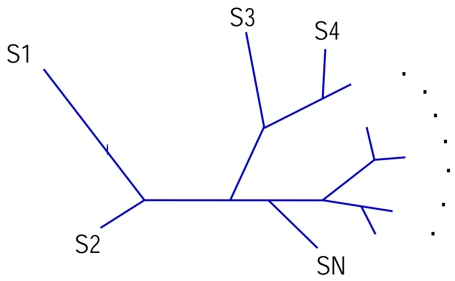
and

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

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

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

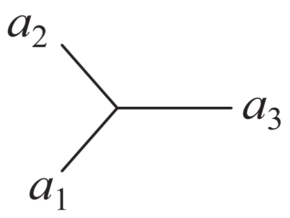
i.e., by all 3×3 minors of all matrix flattenings of P on edges of T.



DNA: $\kappa = 4$ states

Edge invariant construction works for any κ giving $(\kappa + 1) \times (\kappa + 1)$ minors of edge flattenings. But, for $\kappa > 2$, edge invariants cannot generate I_T.

Example: Consider T_3,

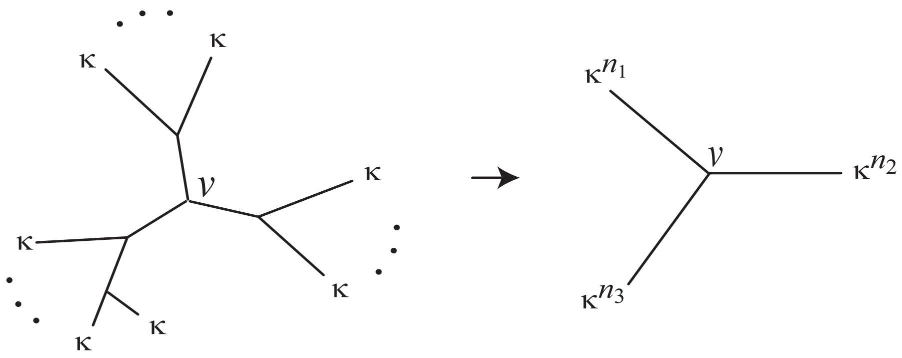


any edge flattening is $\kappa \times \kappa^2$, so no minors of size $(\kappa + 1) \times (\kappa + 1)$, i.e., no edge invariants

But $\dim(V_T) < \kappa^3 - 1$ by counting parameters, so invariants exist.

(The homogeneous degree-5 component of I_T is 1728-dimensional.)

For an arbitrary tree, focus on a node:



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

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

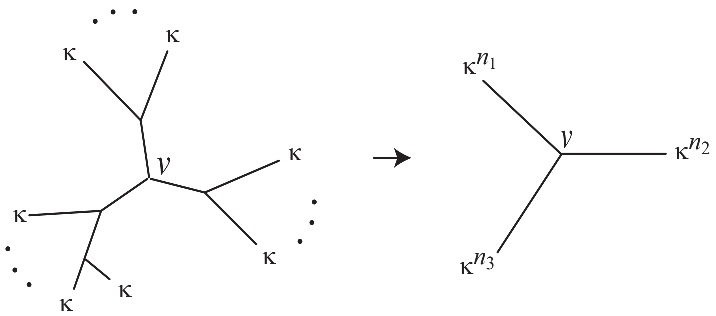
a $\kappa^{n_1} \times \kappa^{n_2} \times \kappa^{n_3}$ tensor, $n_1 + n_2 + n_3 = n$.

Then

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

$$\Big(\text{i.e., } \mathrm{Flat}_v(P) \in \mathrm{Sec}^{\kappa}(\mathbb{P}^{\kappa^{n_1}-1} \times \mathbb{P}^{\kappa^{n_2}-1} \times \mathbb{P}^{\kappa^{n_3}-1}) \Big)$$

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



- a 3-d tensor of the form $\vec{a} \otimes \vec{b} \otimes \vec{c}$ is of rank 1
- a tensor that is a sum of κ rank 1 tensors (and no fewer) is of rank κ.

  $\mathrm{Flat}_v(P)$ is the sum of κ rank-1 tensors, one for each possible state at the internal node.

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

- Tensor rank arises naturally to express conditional independence.

Main result for $\kappa > 2$:

Theorem: For any κ, 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.

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

(cf., Sturmfels' talk)

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

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

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 invariants, we can potentially say something about data's support for

— a *particular edge* (= split = bipartition of taxa), or

— a *particular node* (= tripartition of taxa)

in a phylogenetic tree. Furthermore,

support for all edges and nodes = support for tree

Identifiability of model parameters.

If $\mathcal{T}_n$ denotes n-leaf tree space and $\mathcal{M}$ any choice of model, then we study the parameterization map(s)

$$\begin{aligned} \phi_{\mathcal{M}} : \bigcup_{T \in \mathcal{T}_n} (T, S_T) &\longrightarrow \mathbb{C}^{\kappa^n} \\ (T, s_T) &\longmapsto P = \phi_{\mathcal{M},T}(s_T) \end{aligned}$$

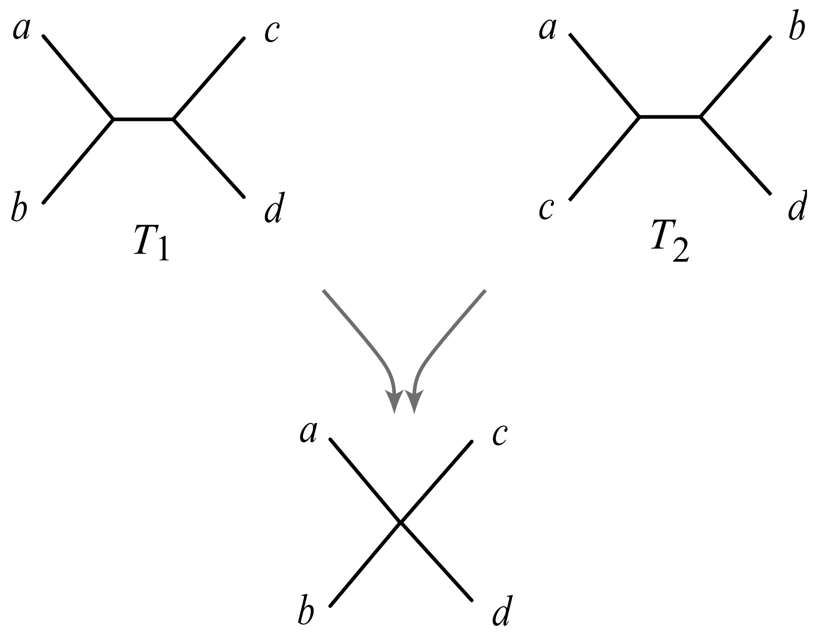
Q. Suppose P is a joint distribution arising from model parameters (T, s_T) for $\mathcal{M}$. Can we *identify* (T, s_T)?

i.e. Is the map $\phi_{\mathcal{M}}$ above invertible?

Identifiability is needed for statistical inference.

Limitations on Identifiability:

For phylogenetics $V_{T_1} \cap V_{T_2} \neq \emptyset$ always (star phylogenies)



Star Phylogeny

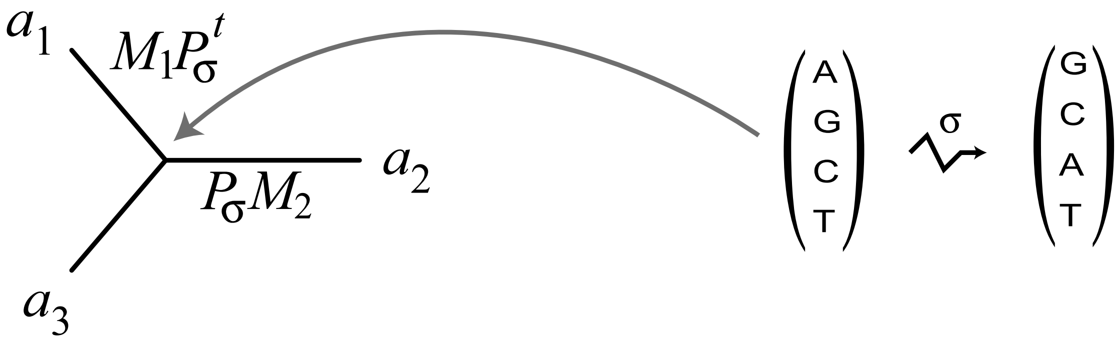
But if $V_{T_1} \cap V_{T_2}$ is a proper subvariety, then the tree is identifiable for generic parameters.

Limitations on Identifiability (cont.):

For numerical parameters,

If states at hidden variables are permuted,

and M_e modified appropriately ⇝ same joint distribution



Refined Q. For a fixed tree T, is ϕ_T generically finite?

If so, what is the cardinality of a generic fiber $\phi_T^{-1}(P)$?

Steel (1994): The tree parameter T is generically identifiable for GM on binary trees T.

Chang (1996): Numerical parameters for GM are generically identifiable up to ‘label swapping’ at internal nodes.

These results also apply to submodels (e.g., group-based)

What about more elaborate models?

Models incorporating additional biological assumptions or complexity.

- Algebraic mixture models:

  $$\text{GM+GM+GM} = \mathrm{Sec}^3(V_T),$$

  (3 rate classes: slow, med, fast)

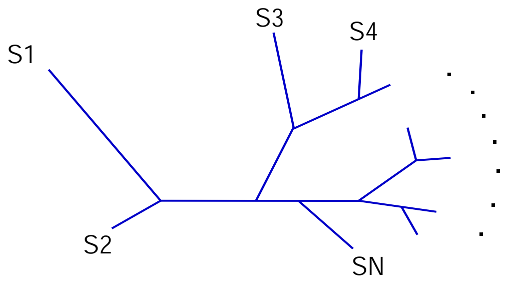
  $$\text{JC+I, GM+I} = \mathrm{Join}(V_T, V_I), \text{ etc.}$$

  (invariable sites, due to functional constraints)

- Continuous-time models: GTR, GTR+I+Γ, Covarion

Note: Continuous-time models are *not algebraic*, but *some* of them can be embedded in algebraic models.

Theorem: Trees are identifiable for generic parameters for a generalized GM model with



$M_{\textit{internal}}$ of size $\lambda \times \lambda$

$M_{\textit{pendant}}$ of size $\lambda \times \kappa$, $\quad \lambda < \kappa^2$

Proof: Construction of only *a few* invariants in I_T, but enough to show $V_T \neq V_{T'}$ if $T \neq T'$.

To obtain results for models of more direct interest, specialize.

Example. GM+GM+GM, $\kappa = 4$:

If on an internal edge e the 3 classes mutate by 4×4 matrices

$$M_{e,1},\ M_{e,2},\ M_{e,3},$$

then let

$$M_e = \begin{pmatrix} M_{e,1} & 0 & 0 \\ 0 & M_{e,2} & 0 \\ 0 & 0 & M_{e,3} \end{pmatrix}.$$

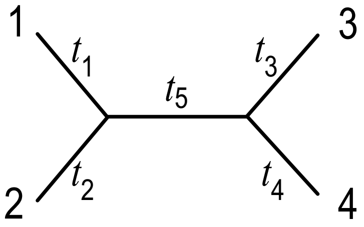
Note $\lambda = 12 < 16 = \kappa^2$, and one can show this is sufficiently generic to apply theorem.

Continuous-time models

General Time-Reversible (GTR) model assumes

- Common rate matrix Q gives the instantaneous rates of various substitutions over all of T.
- Parameter t_e denotes elapsed time along edge e.
- Substitution matrices on edges are $M_e = \exp(Qt_e)$.
- Root distribution $\boldsymbol{\pi}$ is an eigenvector of Q with eigenvalue 0,

  (of M_e with eigenvalue 1).
- $\mathrm{diag}(\pi)Q$ is symmetric.

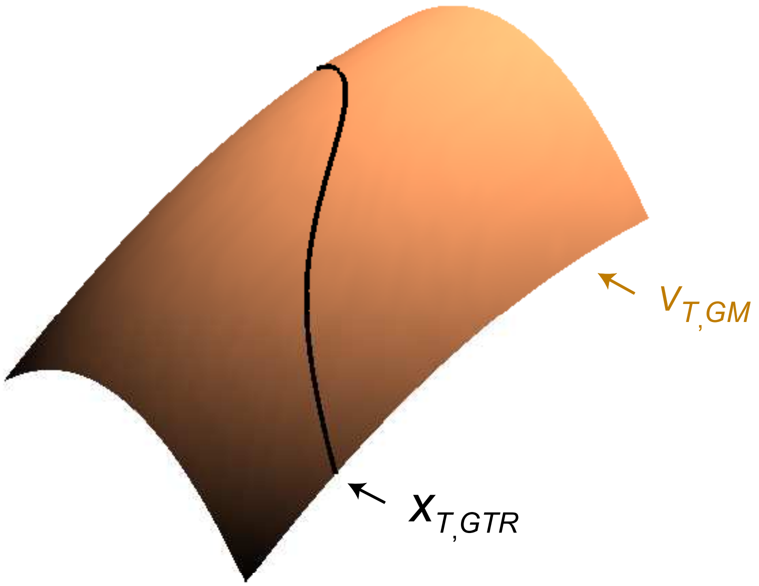


Q = rate matrix

$$\frac{d(A \rightarrow G)}{dt}$$

$M_e = \exp(Qt_e)$

For studying identifiability via $\phi_{\mathcal{M}}$, the matrix exponential puts the model outside of polynomial setting, but GTR embeds in GM



Models commonly used in data analysis (GTR, GTR$+$I$+\Gamma$, covarion)

- continuous-time model is appealing to some (time of descent)
- reduces number of parameters
- extends to rates-across-sites models:

$+\Gamma$: additional parameter λ_i for each site drawn
from Γ distribution, $M_e = \exp(Q\lambda_i t_e)$ for that site

$+$I: allows some sites to be Invariable

$+$I$+\Gamma$: both.

Note: $+\Gamma$ is a *continuous mixture* $\rightsquigarrow$ no algebraic variety

However,

Finite mixtures embed in algebraic models

Corollary. Tree is identifiable for GTR+(3 rate-classes)

For the most commonly-used model — GTR+I+Γ —

it has not been proved that the tree parameter is identifiable.

History:

—— (1990s): GTR is identifiable

Rogers (2001): flawed proof for GTR+I+Γ

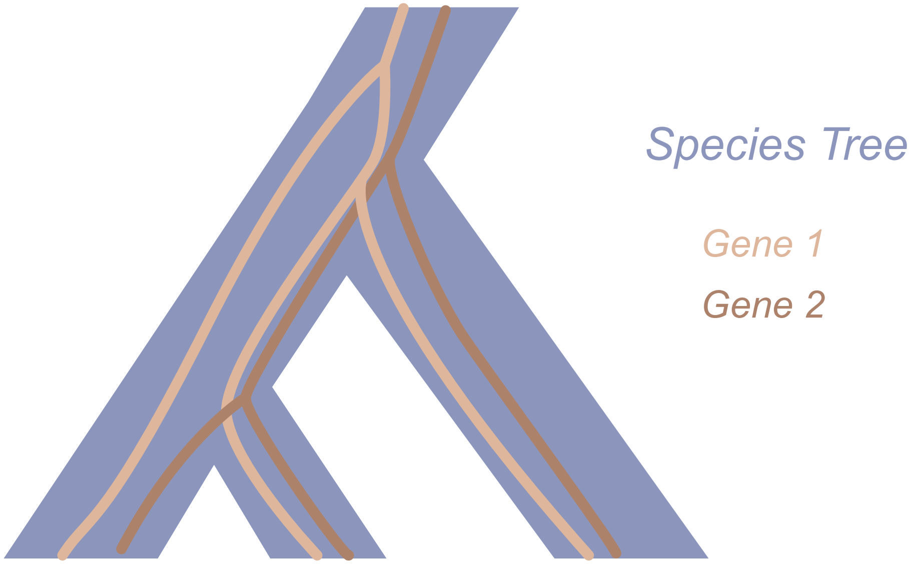
A-Rhodes (2004-7): GTR+I

A-Ané-Rhodes (2007): GTR+Γ (no variety)

Tree mixtures.

Different parts of sequences may have evolved along different trees

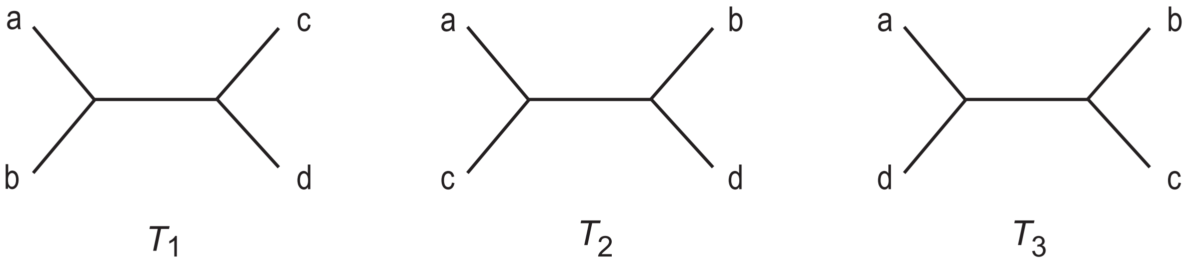
— gene tree vs. species tree, incomplete lineage sorting



— horizontal gene transfer

Simple model:

4-taxon trees T_1, T_2, T_3



Joint distributions $P_{1,2}$ are two-tree mixtures

$$P_{1,2} = \delta P_{T_1,\mathcal{M}} + (1-\delta)P_{T_2,\mathcal{M}}$$

with δ a mixing parameter.

$$V_{T_1,T_2} = \mathrm{Join}(V_{T_1}, V_{T_2})$$

Theorem. Suppose P is a joint distribution arising from a 2-tree GM mixture on 4-taxon trees for $\kappa = 4$ states. Then the trees T_i, T_j and stochastic parameters s_i, s_j are generically locally identifiable.

Similarly for 2-tree GTR mixtures.