On the identifiability of two tree mixtures for group-based models

E. Allman[1] S. Petrović[2] J. Rhodes[1]
S. Sullivant[3]

[1] University of Fairbanks, Alaska
[2] University of Illinois Chicago
[3] North Carolina State University

Phylomania 2010 – Hobart, Tasmania
November 2010

Today's talk:

- identifiability of 2-tree mixture models
- work dates from 2009 and before
- focus today on algebraic techniques (technical at times)

Background

Interest sparked by papers/conversations

- Kolaczkowski and Thornton: 2004 Nature
- Mossel and Vigoda; Ronquist et al: 2005, 2006, Science
- Štefankovič and Vigoda: 2007, JCB, Phylogeny of Mixture Models: Robustness of Maximum Likelihood and Non-identifiable Distributions
- Matsen and Steel: 2007, Sys. Bio., Phylogenetic mixtures on a single tree can mimic a tree of another topology
- Matsen, Mossel, and Steel: 2008, BMB, Mixed-up trees: the structure of phylogenetic mixtures
- Junhyong Kim

Due to incomplete lineage sorting, or other biological phenonomenon, sequence data may have evolved along two or more trees.

Q: Is it theoretically possible to identify the two trees giving rise to expected pattern frequencies?

Q': If so, what about the numerical parameters for these trees?

Due to incomplete lineage sorting, or other biological phenonomenon, sequence data may have evolved along two or more trees.

Q: Is it theoretically possible to identify the two trees giving rise to expected pattern frequencies?

Q': If so, what about the numerical parameters for these trees?

Modeling sequence evolution along a tree(s)

For a fixed tree T and a model of sequence evolution (GTR, GTR+Γ, JC, ...), the distribution of states at the leaves of T is a function ψ_T of the model's parameters.

Eg. GTR model on a n-taxon tree T

parameterization map

$$\begin{aligned} \psi_T : \quad S_T \quad &\longrightarrow \Delta^{4^n-1} \\ \left(\pi, Q, \{t_e\}\right) &\longmapsto P = (p_{i_1 \cdots, i_n}) \end{aligned}$$

where $p_{i_1 \cdots i_n}$ is the expected frequency of pattern $\mathbf{i} = i_1 \cdots i_n$ at the leaves of T.

Mixture models

Modeling sequence evolution along two or more trees requires using a *mixture model*.

Eg. Suppose T_1 and T_2 are two n-taxon trees, then the distribution is a point in the image of

$$\begin{aligned} \psi_{T_1,T_2} : \; S_{T_1} \times S_{T_2} \times [0,1] &\longrightarrow \Delta^{4^n-1} \\ (s_1, s_2, w) &\longmapsto P = (p_{i_1 \cdots, i_n}) \end{aligned}$$

where

$$P = w\psi_{T_1}(s_1) + (1-w)\psi_{T_2}(s_2)$$

is the weighted sum of the distributions for parameter choices on T_1 and T_2.

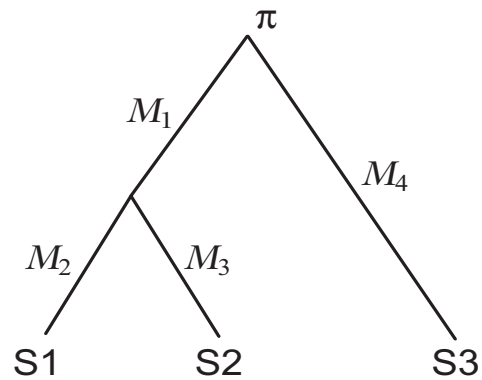
Group-based models

Today: focus on *group-based models*

Cavender-Farris-Neyman (CFN), Jukes-Cantor (JC),
Kimura 2-Parameter (K2P), Kimura 3-Parameter (K3P)

These models, as well as GM, have an *algebraic* structure useful for analysis.

Model parameters π, $\{M_e\}$ on tree T



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

lead to a *polynomial* parameterization map ψ_T.

Thus, any mixture distribution $P_{T_1,T_2} \in \psi_{T_1,T_2}$ is also parameterized by polynomials.

Mixture varieties

Extending the parameterization to complex parameters, define

$$V_{T_1} * V_{T_2} = \overline{\mathrm{Im}\, \psi_{T_1, T_2}},$$

the *phylogenetic mixture variety*.

(Point: This allows ideas from algebraic geometry to be used.)

Algebraic geometry reminders

- Fundamental correspondence:

$$\text{Geometry} \longleftrightarrow \text{Algebra}$$
$$V \longleftrightarrow \mathcal{I}_V$$

  Corresponding to any phylogenetic variety V is its ideal $\mathcal{I}_V$ of *phylogenetic invariants*, the ideal of polynomials f in the pattern frequencies $p_{\mathbf{i}}$ so that

$$f(P) = 0 \text{ for any } P \in V.$$

- Inclusion reversing correspondence:

$$V_1 \subseteq V_2 \iff \mathcal{I}_{V_2} \subseteq \mathcal{I}_{V_1}$$

More notation

For stochastic parameter choices, denote the collection of joint distributions by

$$\mathcal{M}_{T_1} * \mathcal{M}_{T_2}.$$

Note that $\mathcal{M}_{T_1} * \mathcal{M}_{T_2} \subsetneq V_{T_1} * V_{T_2}$.

Though the varieties are used for proofs because of their algebraic structure (dim, good intersection properties, etc.),

all results today hold for the stochastic distributions.

Monomial parameterization

Hendy, Penny, Székely, Erdös, Evans, Speed, Sturmfels, Sullivant:

Group-based models can be diagonalized by means of the discrete Fourier transform over G (Hadamard transform). In the Fourier coordinates, group-based models give rise to toric varieties.

(In this setting, ψ_T is parameterized by monomials.)

Moreover, the discrete Fourier transform is a *linear* change of variables, so it behaves well with respect to taking mixtures of group-based models.

$$\mathcal{F}(\mathcal{M}_{T_1}) * \mathcal{F}(\mathcal{M}_{T_2}) = \mathcal{F}(\mathcal{M}_{T_1} * \mathcal{M}_{T_2})$$

Fourier coordinates

For each split $A|B$ in T, introduce a set of Fourier parameters

$$\{a_g^{A|B} : g \in G\}.$$

Theorem (Hendy-Penny)

In the Fourier coordinates, a group-based phylogenetic model is given parameterically by:

$$q_{g_1,\ldots,g_n} = \begin{cases} \prod_{A|B \in \Sigma(T)} a^{A|B}_{\sum_{a \in A} g_a} & \text{if } g_1 + \cdots + g_n = 0 \\ 0 & \text{if } g_1 + \cdots + g_n \neq 0 \end{cases}$$

'Coordinates' in this parameterization are called *q-coordinates*.

Fourier coordinates

For JC, K2P, we take $G = \mathbb{Z}_2 \times \mathbb{Z}_2 = \{A, C, G, T\}$.

- For K2P model, we have $a_G^{A|B} = a_T^{A|B}$ for all $A|B$
- For JC model, we have $a_C^{A|B} = a_G^{A|B} = a_T^{A|B}$ for all $A|B$.

Tree parameter identifiability (stochastic version)

Definition

The tree parameters $T_1, \ldots, T_k$ in a k-class phylogenetic mixture model are *identifiable*, if for all

$$P \in \mathcal{M}_{T_1} * \cdots * \mathcal{M}_{T_k}$$

there does not exist another set of k trees $T_1', \ldots, T_k'$ such that

$$P \in \mathcal{M}_{T_1'} * \cdots * \mathcal{M}_{T_k'}.$$

Tree parameter identifiability (geometric version)

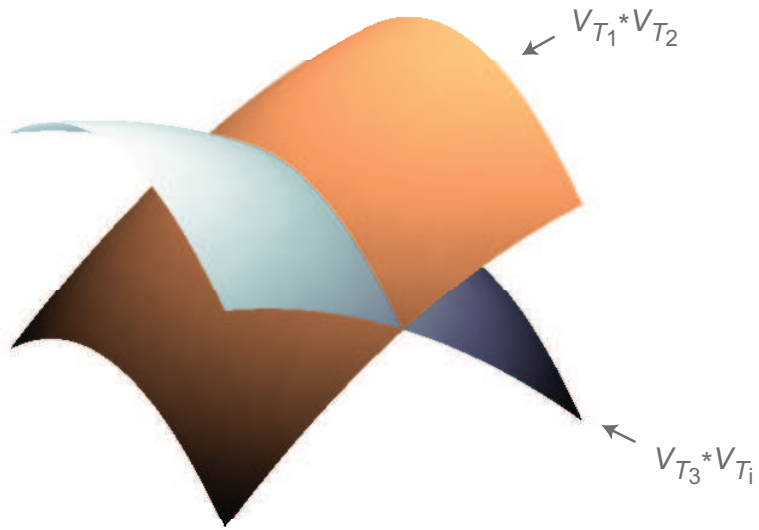
Definition

The tree parameters in a k-class phylogenetic mixture model are

generically identifiable

if for all non-equal multisets $\{T_1, \ldots, T_k\}$, and $\{T'_1, \ldots, T'_k\}$,

$$\dim(V_{T_1} * \cdots * V_{T_k} \cap V_{T'_1} * \cdots * V_{T'_k}) < \dim(V_{T_1} * \cdots * V_{T_k}).$$



Generic identifiability of tree parameters

An immediate consequence of the *geometric* definition:

$$\dim(V_{T_1} * V_{T_2} \cap V_{T_1'} * V_{T_2'}) < \dim(V_{T_1} * V_{T_2})$$

is that tree parameters are generically identifiable for stochastic parameter choices too.

That is, the trees giving rise to

$$\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$$

are identifiable, except on a *non-generic set* $\mathcal{E}$ of stochastic parameters (s_1, s_2, π) of *Lebesque measure zero* where $\psi_{T_1,T_2}(s_1, s_2, \pi) = \psi_{T_1',T_2'}(s_1', s_2', \pi')$.

($\mathcal{E}$ is the set of bad parameters.)

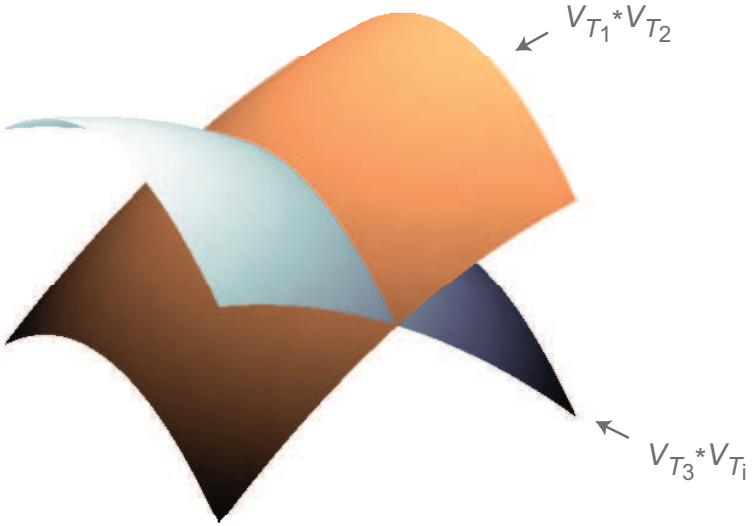
Algebraic methods for proofs

Use

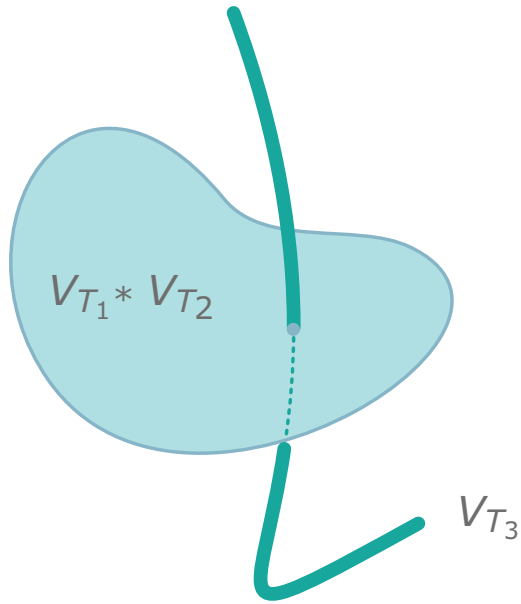
- dimension counts for phylogenetic varieties
- all phylogenetic mixture varieties are irreducible, since they are parameterized
- two irreducible varieties of the same dimension either coincide or intersect in a sub-variety of lower dimension

  Analogy with linear spaces.

  $\Longrightarrow$ if two phylogenetic varieties are distinct, then parameters will be generically identifiable
- two varieties V_1 and V_2 are distinct if $\mathcal{I}_{V_1} \neq \mathcal{I}_{V_2}$ and $V_1 \nsubseteq V_2$ if there exists an invariant $f_2 \in \mathcal{I}_{V_2} \setminus \mathcal{I}_{V_1}$



$$\mathcal{I}_{V_{T_1} * V_{T_2}} \neq \mathcal{I}_{V_{T_3} * V_{T_4}}$$



$$\exists f \in \mathcal{I}_{V_{T_1} * V_{T_2}} \setminus \mathcal{I}_{V_{T_3}}$$

Algebraic methods for proofs

Use

- group-based models (JC and K2P) have linear invariants which can be used to construct invariants for 2-tree mixtures
- computational algebra packages like `Singular`

Main theorem (tree parameters)

Theorem

The tree parameters of the 2-tree mixture model $\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$ are generically identifiable under the Jukes-Cantor and Kimura 2-parameter models if T_1, T_2 are binary with $n \geq 4$ leaves.

Strategy:

Prove theorem for quartets $n = 4$, then lift to larger trees.

Identifiability of quartet trees

Proposition

Let $T_1 = 12|34, T_2 = 14|23, T_3 = 13|24$. *Then*

$$\ell(q) = q_{GGGG} + q_{GTGT} - q_{GGTT} - q_{GTTG}$$

satisfies $\ell(q) = 0$ *for all* $q \in \mathcal{M}_{T_1} * \mathcal{M}_{T_2}$,

but $\ell(q) \neq 0$ *for some* $q \in \mathcal{M}_{T_3}$ *for the JC and K2P models.*

Corollary

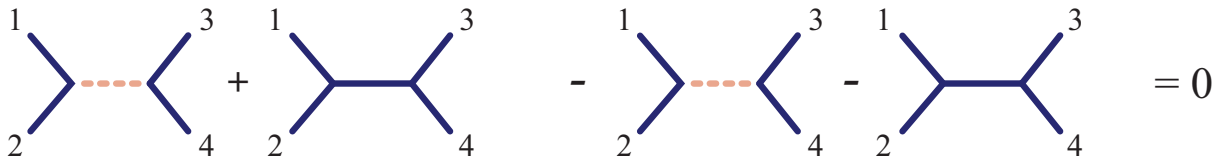
Generic identifiabiliity of tree parameters holds for $n = 4$.

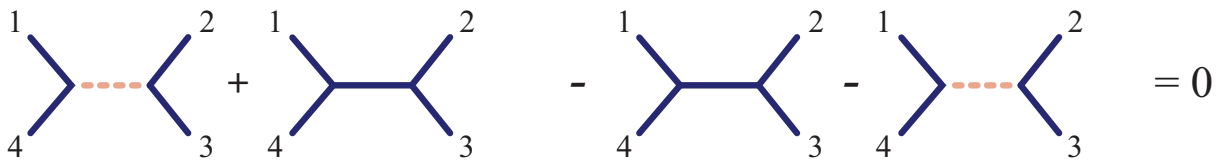
a few details ....

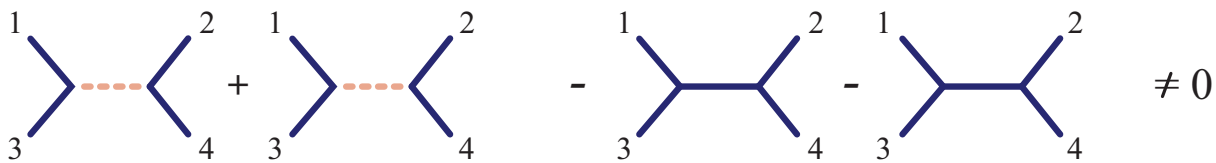
If $q = wq^1 + (1-w)q^2$, then since ℓ is linear

$$\implies \ell(q) = \ell(wq^1 + (1-w)q^2) = w\ell(q^1) + (1-w)\ell(q^2)$$

$\ell(q) = q_{GGGG} + q_{GTGT} - q_{GGTT} - q_{GTTG}$

T_1 

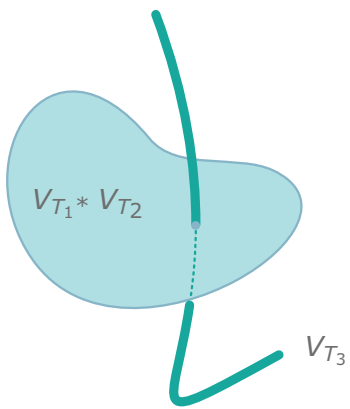
T_2 

T_3 

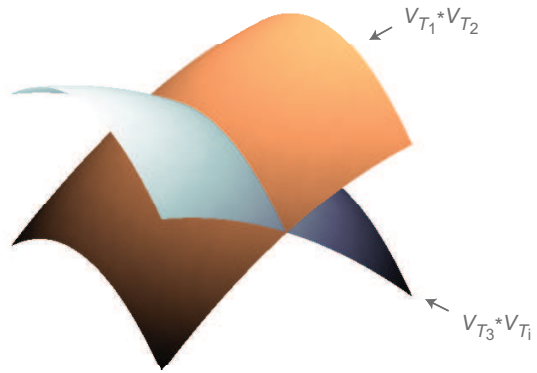
Generic identifiabiliity of tree parameters holds for $n = 4$.

Case: two different trees in mixture

- The linear invariant $\ell \in \mathcal{I}_{V_{T_1} * V_{T_2}} \setminus \mathcal{I}_{V_{T_3}}$. Thus, $V_{T_3} \nsubseteq V_{T_1} * V_{T_2}$.
- Since $V_{T_3} \subset V_{T_3} * V_{T_i}$, we have $\mathcal{I}_{V_{T_3} * V_{T_i}} \subset \mathcal{I}_{V_{T_3}}$ and thus, $\ell \in \mathcal{I}_{V_{T_1} * V_{T_2}} \setminus \mathcal{I}_{V_{T_3} * V_{T_i}}$.
- Since $V_{T_1} * V_{T_2}$ and $V_{T_3} * V_{T_i}$ are irreducible of the same dimension with different ideals, they are distinct.



$$\exists \ell \in \mathcal{I}_{V_{T_1} * V_{T_2}} \setminus \mathcal{I}_{V_{T_3}}$$



$$\mathcal{I}_{V_{T_1} * V_{T_2}} \neq \mathcal{I}_{V_{T_3} * V_{T_i}}$$

Generic identifiability of continuous parameters

Definition

The continuous parameters of a 2-tree mixture model are

generically identifiable

if for generic choices of (s_1, s_2, w),

$$\psi_{T_1,T_2}(s_1, s_2, w) = \psi_{T_1,T_2}(s_1', s_2', w')$$

implies

$$(s_1, s_2, w) = (s_1', s_2', w')$$

or, in the case where $T_1 = T_2$, that

$(s_1, s_2, w) = (s_2', s_1', 1 - w')$.

Main theorem (continuous parameters)

Theorem*

The continuous parameters of the 2-tree mixture model $\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$ are generically identifiable under the Jukes-Cantor and Kimura 2-parameter models if T_1, T_2 are binary with $n \geq 5$ leaves.

Definition

Theorem* means that the result holds *with high probability*.

Note: If $T_1 = T_2$, no * needed.

Theorem* ?

Proposition: *Let $\psi : \mathbb{C}^d \to \mathbb{C}^m$ be a polynomial (or rational) map. Then for some $k \in \{1, 2, 3, \ldots, \infty\}$,*

$$\psi \text{ is generically } k-\text{to}-1.$$

That is, except for some exceptional set $\mathcal{E}$ of parameter space, the map will be $k-$to-1. Moreover, $\mathcal{E}$ is of Lebesgue measure 0 within the parameter space.

For example, $\psi : \mathbb{C} \to \mathbb{C}$ given by

$$\psi(z) = z^2 \quad \text{and} \quad \psi(z) = \frac{1}{z^2}$$

Taking $\mathcal{E} = \{0\}$, then $k = 2$.

Theorem* ?

Proposition: *Let $\psi : \mathbb{C}^d \to \mathbb{C}^m$ be a polynomial (or rational) map. Then for some $k \in \{1, 2, 3, \ldots, \infty\}$,*

$$\psi \text{ is generically } k-\text{to}-1.$$

That is, except for some exceptional set $\mathcal{E}$ of parameter space, the map will be $k-$to-1. Moreover, $\mathcal{E}$ is of Lebesgue measure 0 within the parameter space.

For example, $\psi : \mathbb{C} \to \mathbb{C}$ given by

$$\psi(z) = z^2 \quad \text{and} \quad \psi(z) = \frac{1}{z^2}$$

Taking $\mathcal{E} = \{0\}$, then $k = 2$.

Polynomial maps are generically $k-$ to -1

1. To prove* the Theorem* for a particular tree, repeatedly generate *random rational parameter choices* θ and then symbolically solve the simultaneous polynomial system

$$\psi(t) = \psi(\theta)$$

   and hope for one solution.

   (One solution means that parameters are 'probably' generically identifiable.)

2. We check this using software `Singular`, for JC and K2P on 4 and 5-taxon trees.

3. Recovering parameters uniquely on quartets $\Longrightarrow$ recover parameters on arbitrarily sized trees.

Why $n = 5$ in Theorem*?

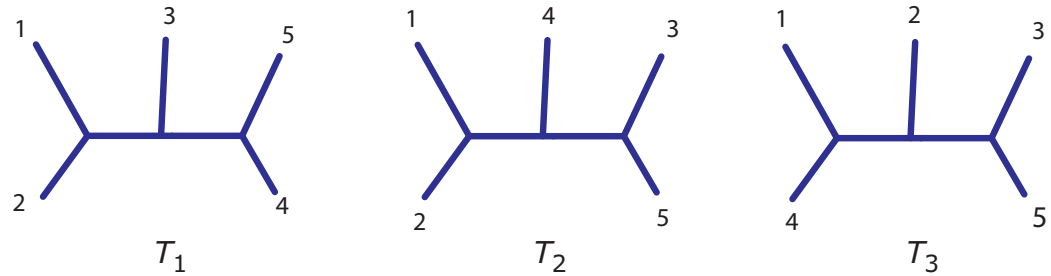
Proposition*

For T a 4-taxon tree under the Jukes-Cantor model, the continuous parameters in $\mathcal{M}_T * \mathcal{M}_T$ are not generically identifiable. The map $\psi_{T,T}$ is generically 6-to-1 (up to label swapping).

For biologically relevant parameters, we observed

between 1 and 4 biologically relevant preimages.

Another Mathematical Surprise



Theorem

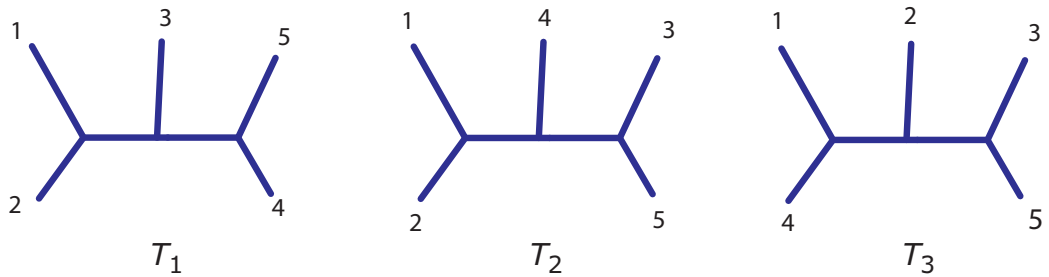
For the Jukes-Cantor model

$$V_{T_2} \subseteq V_{T_1} * V_{T_3}.$$

Q: Is $\mathcal{M}_2 \subseteq \mathcal{M}_1 * \mathcal{M}_3$?

A*: No, unless you allow 0 and/or infinite branch lengths in T_1 and T_3.

Another Mathematical Surprise



Theorem

For the Jukes-Cantor model

$$V_{T_2} \subseteq V_{T_1} * V_{T_3}.$$

Q: Is $\mathcal{M}_2 \subseteq \mathcal{M}_1 * \mathcal{M}_3$?

A*: No, unless you allow 0 and/or infinite branch lengths in T_1 and T_3.

Mixtures of many trees

Recent work of Rhodes and Sullivant has advanced these results:

Theorem

- Under the general Markov model of sequence evolution, the tree parameter and continuous parameters are generically identifiable for a k-class mixture on the same tree, provided

$$k < 4^{\lceil n/4 \rceil - 1}.$$

Open problems

- Develop methods to remove * from Theorem*
- Beyond group-based models: GTR, rate variation
- Arbitrary k-tree mixtures

Thank you.