Identifiability of 3-Class Jukes-Cantor Mixtures

Colby Long and Seth Sullivant

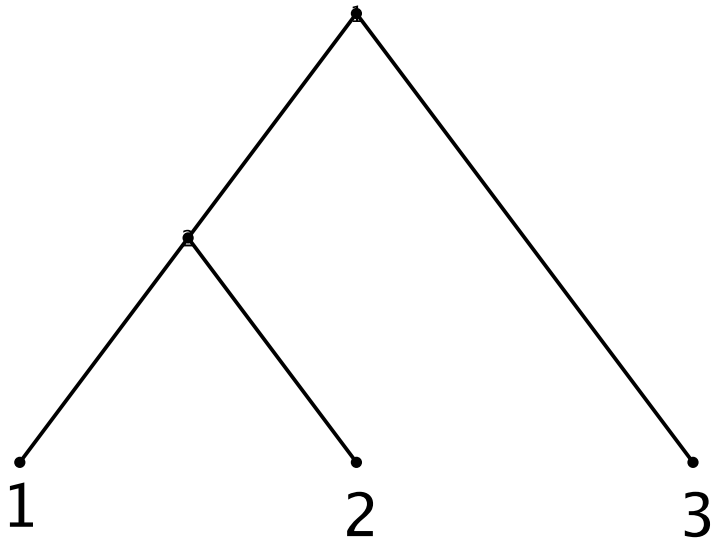
North Carolina State University

October 26, 2014

Phylogenetic Models

Problem

Find a tree that represents the evolutionary history of a group of taxa.



DATA

Species 1: ACCGTAGATGACT...
Species 2: ACTGTAGATGACT...
Species 3: ACCGTACATGACT...

- Latent variable graphical models
- Give probability distribution on n-tuples of DNA characters
- To infer phylogeny we require identifiable parameters

Inferring Phylogeny

Definition

A model parameter is identifiable if the distribution arising from the model uniquely determines the parameter.

Identifiability results have been established for

- Basic models of character evolution (Chang, 1996)
- Covarion and mixture models (Allman and Rhodes)
- Mixture models with various restrictions (Allman, Matsen, Rhodes, Steel, Sullivant)
- Two-class mixtures of group-based models (Allman, Petrovic, Rhodes, Sullivant 2011)

Theorem (L-Sullivant 2014)

The tree parameters of the 3-class Jukes-Cantor mixture model are generically identifiable on trees with ≥ 6 *leaves.*

The Jukes-Cantor Model of DNA Evolution

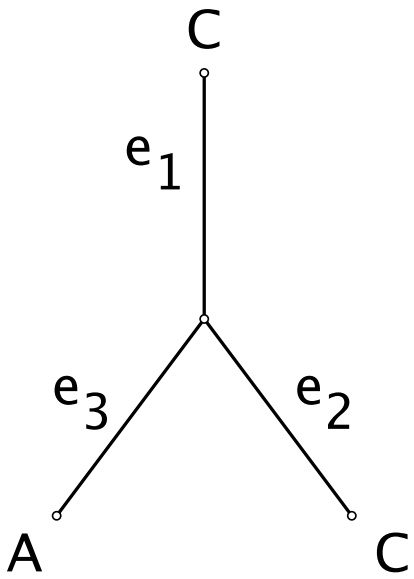
- Tree parameter: Binary leaf-labelled tree T with label set $[n]$
- Random variable X_v associated to each node of T
- Transition matrix associated to each edge

$$A^e = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_e & \beta_e & \beta_e & \beta_e \\ \beta_e & \alpha_e & \beta_e & \beta_e \\ \beta_e & \beta_e & \alpha_e & \beta_e \\ \beta_e & \beta_e & \beta_e & \alpha_e \end{pmatrix} \end{array}$$

$$A^e_{ij} = P(X_v = i | X_w = j)$$

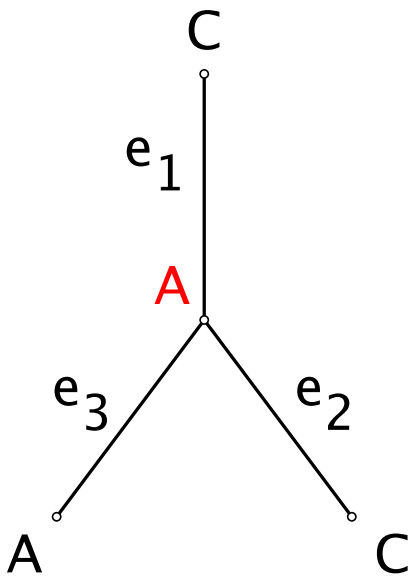
- To find the probability of observing a particular state at the leaves sum over all states of internal nodes.

Jukes-Cantor Example



$$A^e = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_e & \beta_e & \beta_e & \beta_e \\ \beta_e & \alpha_e & \beta_e & \beta_e \\ \beta_e & \beta_e & \alpha_e & \beta_e \\ \beta_e & \beta_e & \beta_e & \alpha_e \end{pmatrix} \end{array}$$

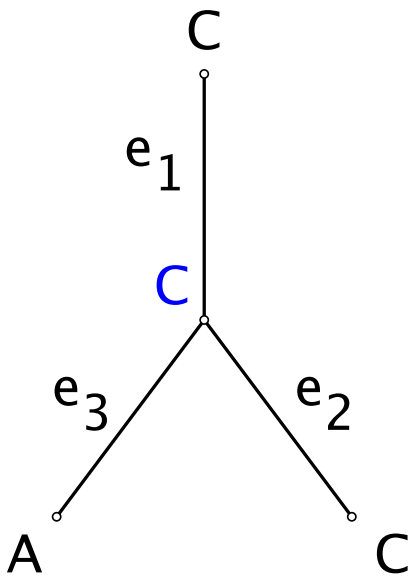
Jukes-Cantor Example



$$A^e = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_e & \beta_e & \beta_e & \beta_e \\ \beta_e & \alpha_e & \beta_e & \beta_e \\ \beta_e & \beta_e & \alpha_e & \beta_e \\ \beta_e & \beta_e & \beta_e & \alpha_e \end{pmatrix} \end{array}$$

$$p_{CCA} = \beta_1\beta_2\alpha_3 +$$

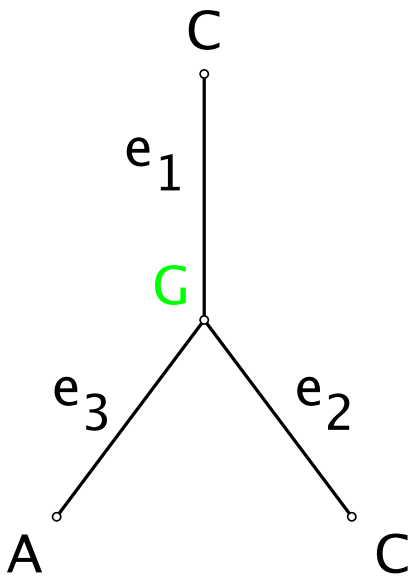
Jukes-Cantor Example



$$A^e = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_e & \beta_e & \beta_e & \beta_e \\ \beta_e & \alpha_e & \beta_e & \beta_e \\ \beta_e & \beta_e & \alpha_e & \beta_e \\ \beta_e & \beta_e & \beta_e & \alpha_e \end{pmatrix} \end{array}$$

$$p_{CCA} = \beta_1\beta_2\alpha_3 + \alpha_1\alpha_2\beta_3 +$$

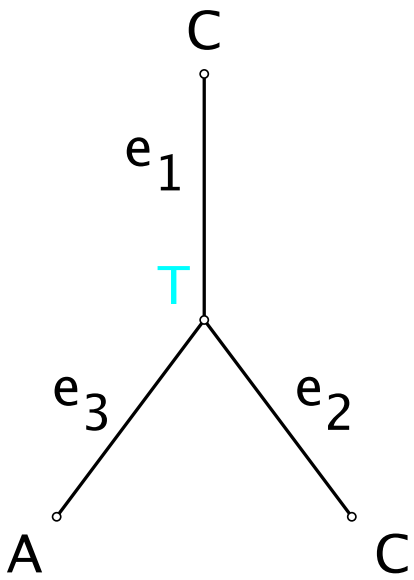
Jukes-Cantor Example



$$A^e = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_e & \beta_e & \beta_e & \beta_e \\ \beta_e & \alpha_e & \beta_e & \beta_e \\ \beta_e & \beta_e & \alpha_e & \beta_e \\ \beta_e & \beta_e & \beta_e & \alpha_e \end{pmatrix} \end{array}$$

$$p_{CCA} = \beta_1\beta_2\alpha_3 + \alpha_1\alpha_2\beta_3 + \beta_1\beta_2\beta_3 +$$

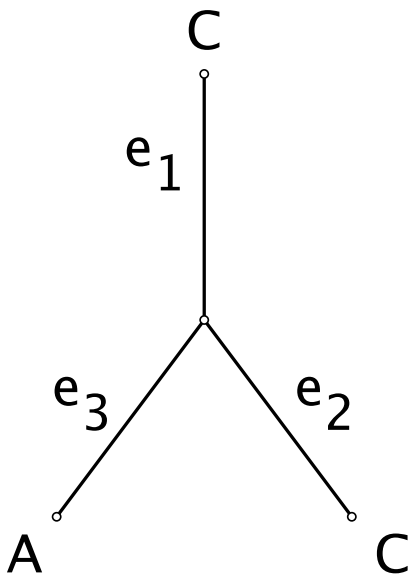
Jukes-Cantor Example



$$A^e = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_e & \beta_e & \beta_e & \beta_e \\ \beta_e & \alpha_e & \beta_e & \beta_e \\ \beta_e & \beta_e & \alpha_e & \beta_e \\ \beta_e & \beta_e & \beta_e & \alpha_e \end{pmatrix} \end{array}$$

$$p_{CCA} = \beta_1\beta_2\alpha_3 + \alpha_1\alpha_2\beta_3 + \beta_1\beta_2\beta_3 + \beta_1\beta_2\beta_3$$

Jukes-Cantor Example



$$A^e = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_e & \beta_e & \beta_e & \beta_e \\ \beta_e & \alpha_e & \beta_e & \beta_e \\ \beta_e & \beta_e & \alpha_e & \beta_e \\ \beta_e & \beta_e & \beta_e & \alpha_e \end{pmatrix} \end{array}$$

$$p_{CCA} = \beta_1\beta_2\alpha_3 + \alpha_1\alpha_2\beta_3 + \beta_1\beta_2\beta_3 + \beta_1\beta_2\beta_3$$

- $\psi_T : \Theta_T \to \Delta^{4^n-1} \subseteq \mathbb{R}^{4^n}$
- $V_T = \overline{\mathrm{im}(\psi_T)}$ is a complex algebraic variety
- $\mathcal{I}(V_T)$ is the ideal of phylogenetic invariants

Mixture Models

Problem

A single tree may not describe our data due to

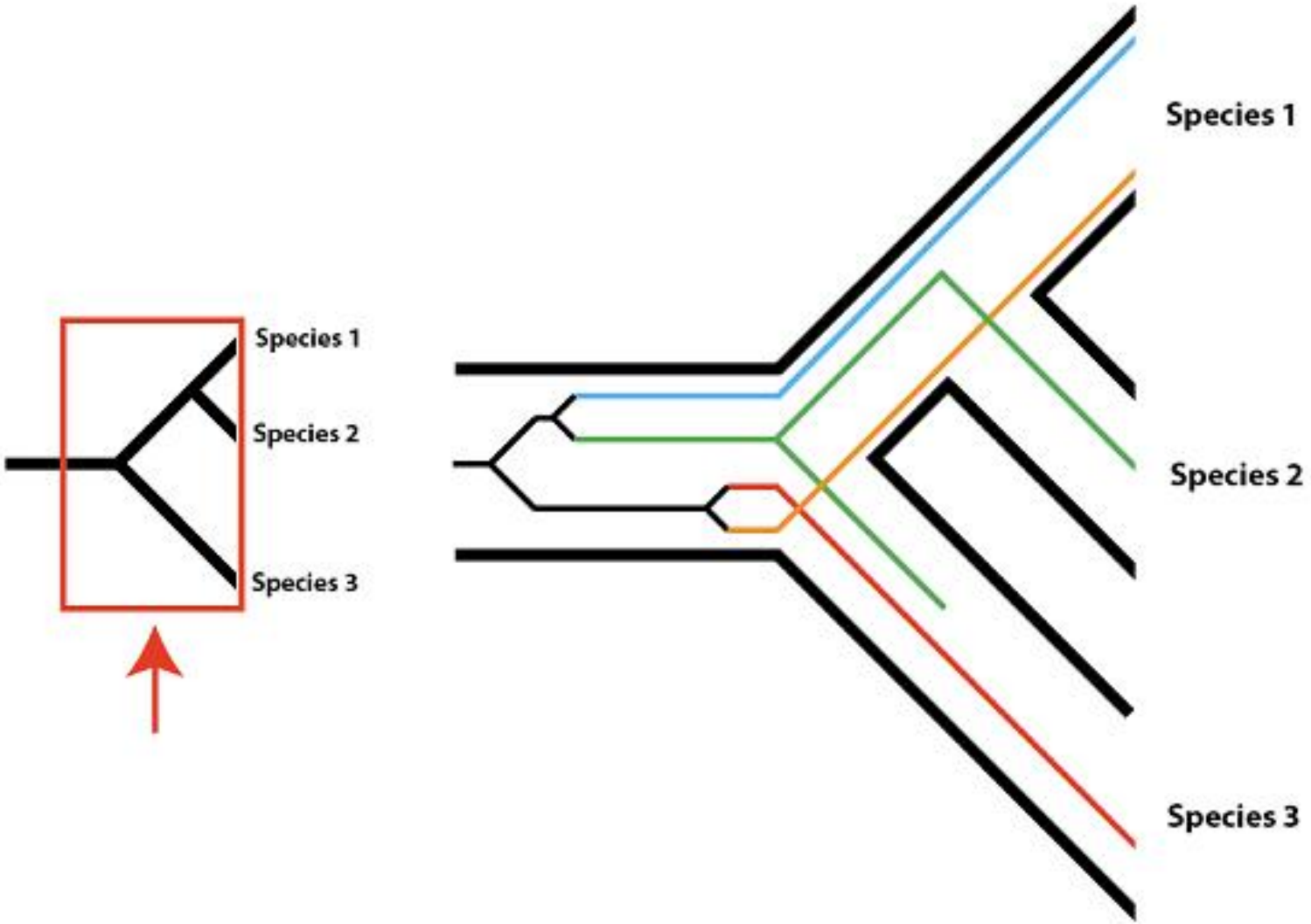
- Horizontal gene transfer
- Incomplete lineage sorting
- Different rates of mutation in the genome

- A mixture model weights the distributions from multiple trees.

$$\psi_{T_1,T_2,T_3} : \Theta_{T_1} \times \Theta_{T_2} \times \Theta_{T_3} \times \Delta^2 \to \Delta^{k^n-1}$$

$$(s_1, s_2, s_3, \pi) \mapsto \pi_1\psi_1(s_1) + \pi_2\psi_2(s_2) + \pi_3\psi_3(s_3)$$

$$\overline{\mathrm{im}(\psi_{T_1,T_2,T_3})} = V_{T_1} * V_{T_2} * V_{T_3}$$


Species 1
Species 2
Species 3
Species 1
Species 2
Species 3

Generic Identifiability of the Tree Parameters

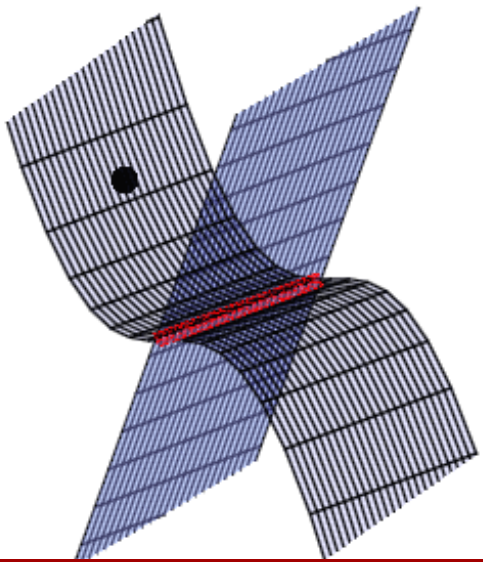
Definition

The tree parameters of an r-tree mixture model are *generically identifiable* for n-leaf trees if for all $S \in \mathcal{T}_{[n],r}$ and generic choices of $(s_1, \ldots, s_r, \pi) \in \Theta_{T_1} \times \cdots \times \Theta_{T_r} \times \Delta^{r-1}$, if there is a $T \in \mathcal{T}_{[n],r}$ and $(s'_1, \ldots, s'_r, \pi') \in \Theta_{T'_1} \times \cdots \times \Theta_{T'_r} \times \Delta^{r-1}$ such that $\psi_T(s_1, \ldots, s_r, \pi) = \psi_{T'}(s'_1, \ldots, s'_r, \pi')$ then $S = T$.

Generic Identifiability of the Tree Parameters

Definition

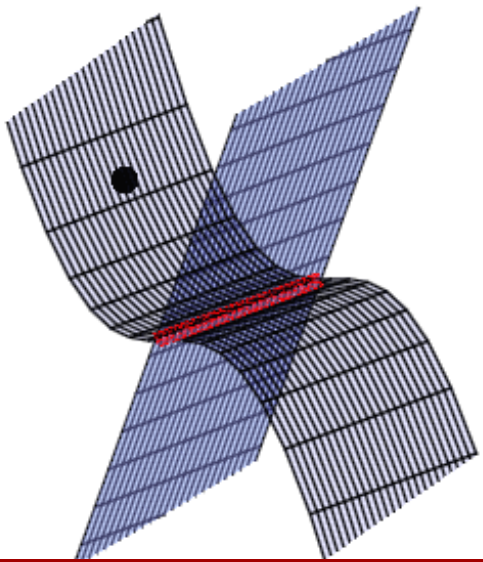
The tree parameters of an r-tree mixture model are *generically identifiable* for n-leaf trees if for all $S \in \mathcal{T}_{[n],r}$ and generic choices of $(s_1, \ldots, s_r, \pi) \in \Theta_{T_1} \times \cdots \times \Theta_{T_r} \times \Delta^{r-1}$, if there is a $T \in \mathcal{T}_{[n],r}$ and $(s'_1, \ldots, s'_r, \pi') \in \Theta_{T'_1} \times \cdots \times \Theta_{T'_r} \times \Delta^{r-1}$ such that $\psi_T(s_1, \ldots, s_r, \pi) = \psi_{T'}(s'_1, \ldots, s'_r, \pi')$ then $S = T$.



Generic Identifiability of the Tree Parameters

Definition

The tree parameters of an r-tree mixture model are *generically identifiable* for n-leaf trees if for all $S \in \mathcal{T}_{[n],r}$ and generic choices of $(s_1, \ldots, s_r, \pi) \in \Theta_{T_1} \times \cdots \times \Theta_{T_r} \times \Delta^{r-1}$, if there is a $T \in \mathcal{T}_{[n],r}$ and $(s'_1, \ldots, s'_r, \pi') \in \Theta_{T'_1} \times \cdots \times \Theta_{T'_r} \times \Delta^{r-1}$ such that $\psi_T(s_1, \ldots, s_r, \pi) = \psi_{T'}(s'_1, \ldots, s'_r, \pi')$ then $S = T$.



- To establish generic identifiability for the tree parameters for 3-tree JC mixtures on n leaves, want to show that for all $S, T \in \mathcal{T}_{[n],3}$, $\dim(V_S \cap V_T) < min\{dim(V_S), dim(V_T)\}$.

Dimension

Theorem (L-Sullivant 2014)

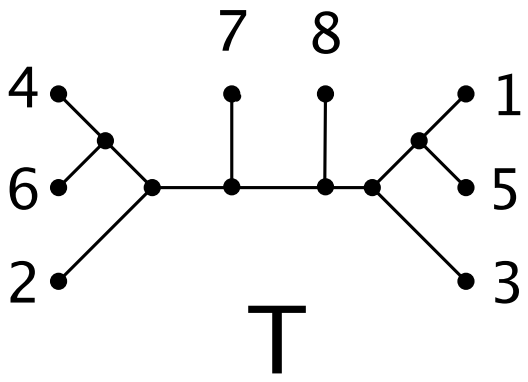
Let $T \in \mathcal{T}_{[n],r}$. For $n \geq 4$ and $r \leq \lceil \frac{n}{2} \rceil$, the join variety $V_{T_1} * \ldots * V_{T_r}$ associated to the r-class Jukes-Cantor mixture model is nondefective.

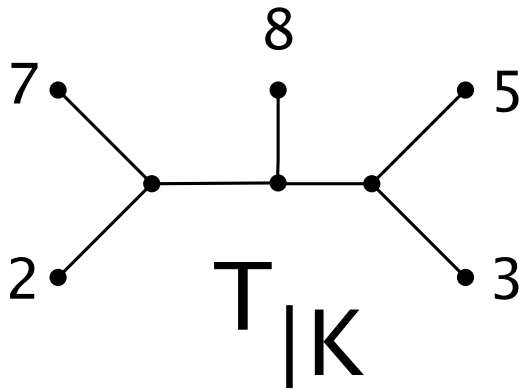
- Proof uses tropical secant dimension approach (Draisma 2008).
- Theorem implies that for all $n \geq 5$ and $S, T \in \mathcal{T}_{[n],3}$, $dim(V_S) = dim(V_T)$.
- Since V_S, V_T are irreducible and $(V_S \cap V_T)$ is contained in both, $\dim(V_S \cap V_T) = min\{dim(V_S), dim(V_T)\} \Rightarrow V_S = (V_S \cap V_T) = V_T$.
- Therefore, it is enough to show that for all distinct $S, T \in \mathcal{T}_{[n],3}$, $V_T \neq V_S$.

$$V_S \neq V_T \iff \mathcal{I}(V_S) \neq \mathcal{I}(V_T)$$

Disentangling Trees

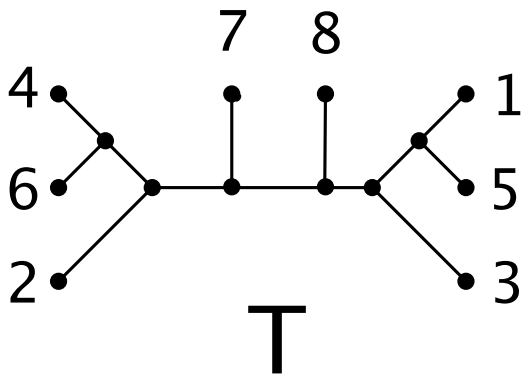
$T \in \mathcal{T}_{[8]}$ $K = \{2, 3, 5, 7, 8\}$

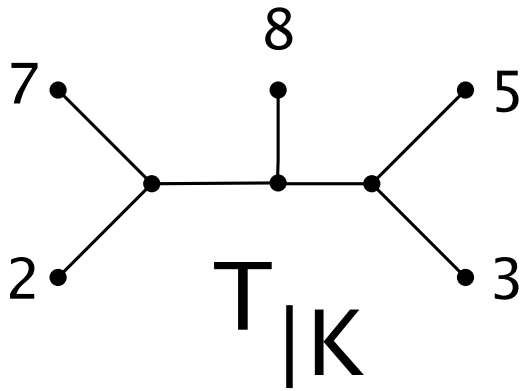




Disentangling Trees

$T \in \mathcal{T}_{[8]}$ $K = \{2, 3, 5, 7, 8\}$

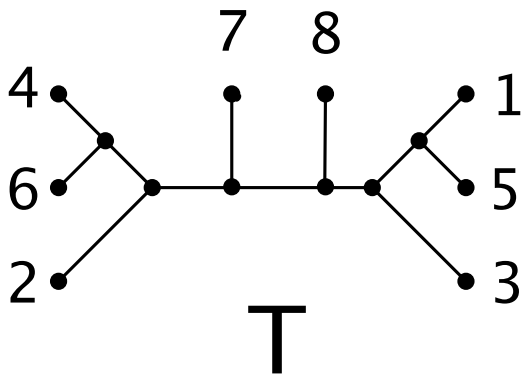


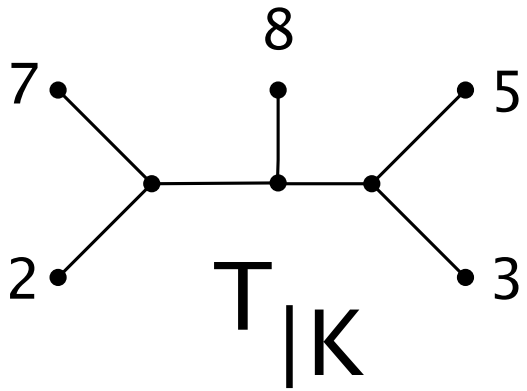


- For $T = \{T_1, T_2, T_3\} \in \mathcal{T}_{[n],3}$, $T_{|K}$ is the multiset $\{T_{1|K}, T_{2|K}, T_{3|K}\}$
- If $S_{|K} \neq T_{|K}$ then $K \subseteq [n]$ is said to disentangle S and T

Disentangling Trees

$T \in \mathcal{T}_{[8]}$ $K = \{2, 3, 5, 7, 8\}$





- For $T = \{T_1, T_2, T_3\} \in \mathcal{T}_{[n],3}$, $T_{|K}$ is the multiset $\{T_{1|K}, T_{2|K}, T_{3|K}\}$
- If $S_{|K} \neq T_{|K}$ then $K \subseteq [n]$ is said to disentangle S and T

Lemma (L-Sullivant 2014)

For all n and distinct $S, T \in \mathcal{T}_{[n],3}$, there exists a disentangling set K with $|K| \leq 6$.

Separating Pairs of Triplets

Lemma

Let $S, T \in \mathcal{T}_{[n],r}$ and $K \subseteq [n]$. If $V_{S_{|K}} \neq V_{T_{|K}}$ then $V_S \neq V_T$.

- $V_{S_{|K}}$ and $V_{T_{|K}}$ are the images of V_S and V_T under the marginal map and $f(U) \neq f(W) \Rightarrow U \neq W$.
- Identifiability for 3-tree mixtures of 6-leaf trees $\Rightarrow$ identifiability for 3-tree mixtures of $n \geq 6$ leaves.
- Therefore, it is enough to show that for all distinct $S, T \in \mathcal{T}_{[6],3}$, $\mathcal{I}(V_T) \neq \mathcal{I}(V_S)$.

We want to separate all 6-leaf triplet pairs by finding a polynomial in the ideal of one that is not in the ideal of the other.

Separating Pairs of Triplets

Lemma

Let $S, T \in \mathcal{T}_{[n],r}$ and $K \subseteq [n]$. If $V_{S_{|K}} \neq V_{T_{|K}}$ then $V_S \neq V_T$.

- $V_{S_{|K}}$ and $V_{T_{|K}}$ are the images of V_S and V_T under the marginal map and $f(U) \neq f(W) \Rightarrow U \neq W$.
- Identifiability for 3-tree mixtures of 6-leaf trees $\Rightarrow$ identifiability for 3-tree mixtures of $n \geq 6$ leaves.
- Therefore, it is enough to show that for all distinct $S, T \in \mathcal{T}_{[6],3}$, $\mathcal{I}(V_T) \neq \mathcal{I}(V_S)$.

We want to separate all 6-leaf triplet pairs by finding a polynomial in the ideal of one that is not in the ideal of the other.

Separating with Linear Invariants

Up to relabeling, there are eighty-five 6-leaf tree triplet pairs with the same set of linear invariants.

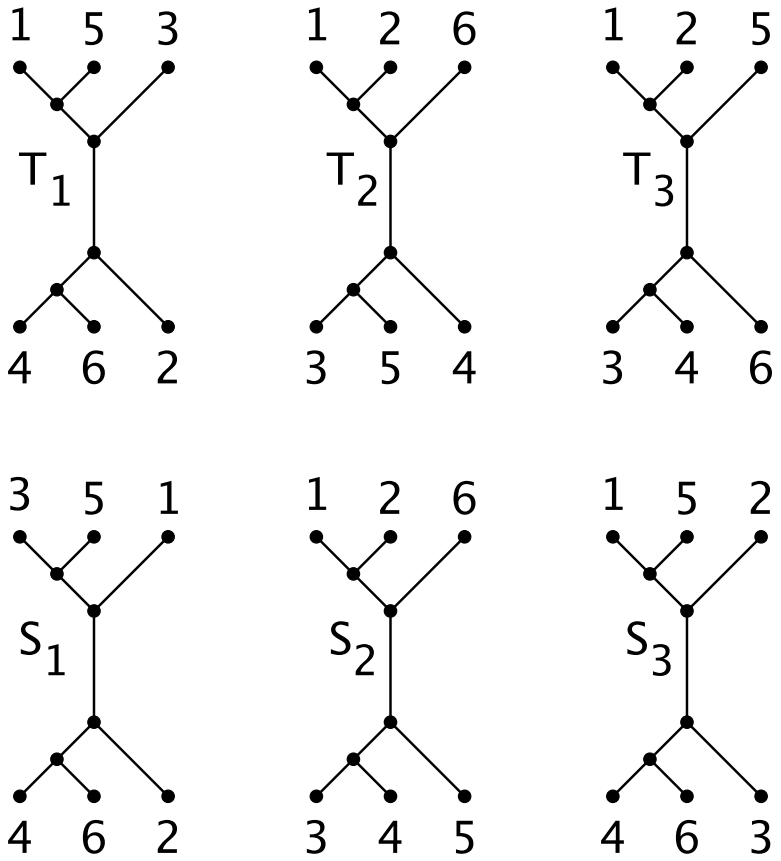
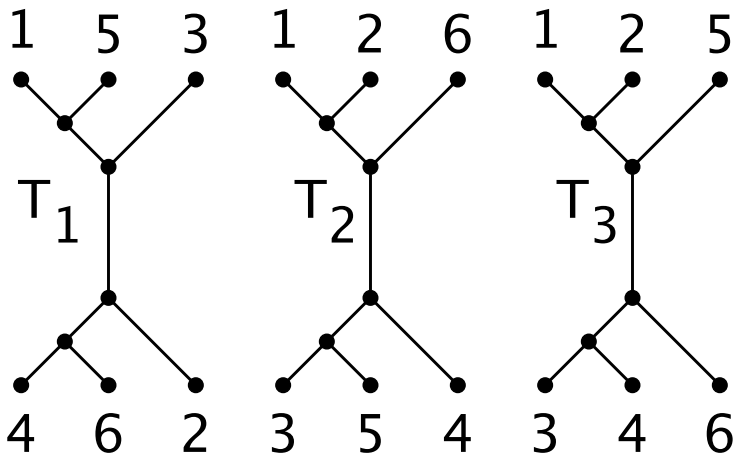


Figure: A 6-leaf triplet pair that is not separated by linear invariants.

Finding Higher Degree Invariants (Example)

- After Fourier/Hadamard Transformation, the parameterization of each V_{T_i} is monomial.

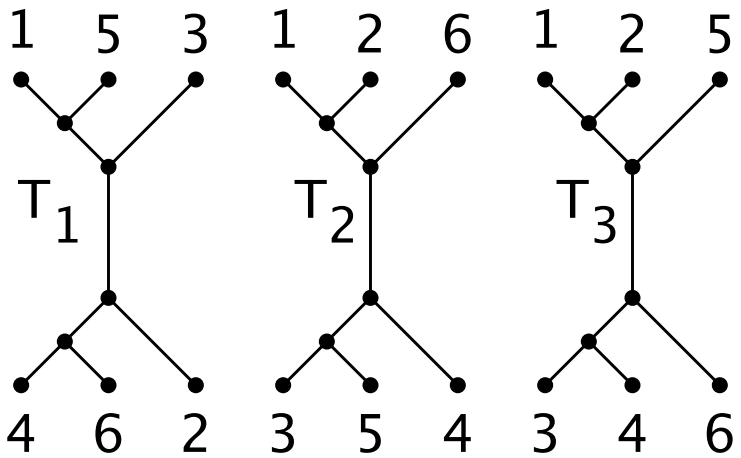


$$q_{CCCAAC} = \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 + \pi_2 b_1 b_2 b_3 b_6 b_8 b_9 + \pi_3 c_1 c_2 c_3 c_6 c_8$$

$$q_{CCGAAG} = \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 a_9 + \pi_2 b_1 b_2 b_3 b_6 b_8 b_9 + \pi_3 c_1 c_2 c_3 c_6 c_8$$

Finding Higher Degree Invariants (Example)

- After Fourier/Hadamard Transformation, the parameterization of each V_{T_i} is monomial.



$$q_{CCCAAC} = \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 + \pi_2 b_1 b_2 b_3 b_6 b_8 b_9 + \pi_3 c_1 c_2 c_3 c_6 c_8$$

$$q_{CCGAAG} = \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 a_9 + \pi_2 b_1 b_2 b_3 b_6 b_8 b_9 + \pi_3 c_1 c_2 c_3 c_6 c_8$$

$\Rightarrow$ $q_{CCCAAC} - q_{CCGAAG} = \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 - \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 a_9$

Finding Higher Degree Invariants (Example)

$$J = \langle y_1 - (\pi_1 a_1 a_2 a_3 a_6 a_7 a_8 - \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 a_9),$$
$$y_2 - (\pi_1 a_1 a_3 a_4 a_5 a_6 a_7 a_8 a_9 - \pi_1 a_1 a_3 a_4 a_5 a_6 a_7), \ldots \rangle.$$

- Eliminate $\{a_1, \ldots, a_9, \pi_1\}$ in J to find $f \in \mathcal{I}(V_T)$.

$$f = (q_{CCCAAC} - q_{CCGAAG})(q_{CACCGT} - q_{CAGGTG}) -$$
$$(q_{CCCAGT} - q_{CCGATC})(q_{CACCAC} - q_{CAGGAC})$$

- Verify $f \notin \mathcal{I}(V_S)$.

Theorem (L-Sullivant 2014)

The tree parameters of the 3-class Jukes-Cantor mixture model are generically identifiable on trees with ≥ 6 leaves.

References


E. Allman, C. Ané, and J.A. Rhodes. Identifiability of a markovian model of a molecular evolution with gamma-distributed rates. *Adv. Appl. Prob.*, 40:229–249, 2008.


E. Allman and J.A. Rhodes. Identifying evolutionary trees and substitution parameters for the general markov model with invariable sites. *Math. Biosci.*, 211(1):18–33, 2008.


E.S. Allman, S. Petrovic, J.A. Rhodes, and S. Sullivant. Identifiability of 2-tree mixtures for group-based models. *IEEE/ACM Trans Comput Biol Bioinformatics*, 8(3):710–722, 2011.


E.S. Allman and J.A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comp. Biol.*, 13(5):1101–1113, 2006.


E.S. Allman, J.A. Rhodes, and S. Sullivant. When do phylogenetic mixture models mimic other phylogenetic models? *Syst. Biol.*, 61(6):1049–1059, 2012.


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


J. Draisma. A tropical approach to secant dimensions. *J. Pure Appl. Algebra*, 212(2):349–363, 2008.


F.A. Matsen, E. Mossel, and M. Steel. Mixed-up trees: the structure of phylogenetic mixtures. *Bull. Math Biol.*, 70(4):1115–1139, 2008.


F.A. Matsen and M. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.*, 56(5):767–775, 2007.


J.A. Rhodes and S. Sullivant. Identifiability of large phylogenetic mixtures. *Bull. Math Biol.*, 74(1):212–231, 2012.


C. Long and S. Sullivant. Identifiability of 3-Class Jukes-Cantor Mixtures. 1406.7256