Identifiability of Phylogenetic Mixture Models

Elizabeth Allman, Sonja Petrović, John Rhodes, and Seth Sullivant

U. of Alaska– Fairbanks, U. of Illinois– Chicago, and NCSU

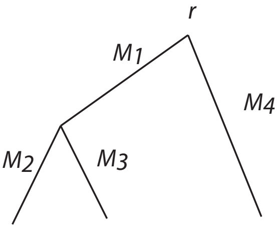
March 27, 2010

Phylogenetic Models

Let T be a trivalent tree with n leaves. Leaves are labeled by $[n] := \{1, 2, 3, \ldots, n\}$.

Associated to each edge of tree e is a Markov (structured) transition matrix M_e.

Once we specify T, and the M_e, get a probability distribution of characters at the leaves of the tree.



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

Think of phylogenetic model as a map

$$\phi_T : \Theta \subseteq \mathbb{R}^k \to \Delta_{4^n}$$

Given by polynomials:
$\mathcal{M}_T := \mathrm{im}\phi_T = \phi_T(\Theta)$, is the phylogenetic model.

Phylogenetic Mixture Models

Suppose there are k classes of sites in the genome.
Each class $j \in [k]$ evolved according to tree T_j on n leaves.
Assuming that the classes are hidden, we observe a probability distribution of the form:

$$\phi_{T_1,\ldots,T_k}(\pi, \{M_e\}) = \pi_1 \cdot \phi_{T_1}(\{M_e^1\}) + \pi_2 \cdot \phi_{T_2}(\{M_e^2\}) + \cdots + \pi_k \cdot \phi_{T_k}(\{M_e^k\})$$

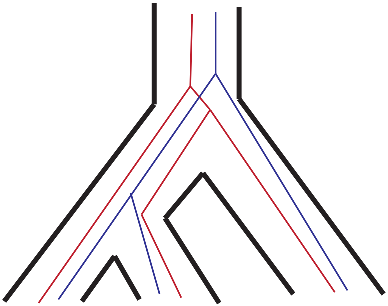
where π_j is the relative proportion of sites of class j.

Definition

Let $T_1, \ldots, T_k$ be trees with n leaves. The phylogenetic mixture model

$$\mathcal{M}_{T_1} * \mathcal{M}_{T_2} * \cdots * \mathcal{M}_{T_k} = \left\{ \sum_{j=1}^{k} \pi_j p^j : \pi_j \geq 0, \sum \pi_j = 1, p^j \in \mathcal{M}_{T_j} \right\}.$$

Why Mixture Models?



- Differing gene tree topologies
- Could explain evolution with recombination

Group-based Models

For remainder we focus on group-based models. Phylogenetic models with structured transition matrices.

$$\begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix} \quad \begin{pmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{pmatrix} \quad \begin{pmatrix} \alpha & \beta & \gamma & \gamma \\ \beta & \alpha & \gamma & \gamma \\ \gamma & \gamma & \alpha & \beta \\ \gamma & \gamma & \beta & \alpha \end{pmatrix} \quad \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha \end{pmatrix}$$

CFN JC K2P K3P

Transition structure is governed by a finite Abelian group G, such that

$$M_e(g,h) = f_e(g-h).$$

Theorem (Evans-Speed 1993, Hendy-Penny 1993)

Group-based models are toric varieties in Fourier coordinates.

The Identifiability Problem

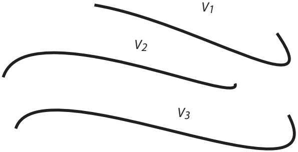
Definition

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

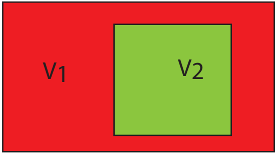
$$p \in \mathcal{M}_{T_1} * \cdots * \mathcal{M}_{T_k}$$

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

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



Identifiable



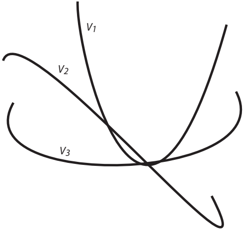
Not Identifiable

Generic Identifiability

Definition

The tree parameters in a k-class phylogenetic mixture model are generically identifiable if for all nonequal multisets $T_1, \ldots, T_k$, and $T'_1, \ldots, T'_k$,

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



Past Work on Identifiability of Tree Mixtures

- Identifiability Results:
  - Allman and Rhodes (2006) $T_1 = \ldots = T_k$, $k < n$.
  - Stefankovic and Vigoda (2007) $T_1 = \ldots = T_k$, JC, K2P
  - Matsen, Mossel, and Steel (2008)
- Non-Identifiability Results:
  - Matsen and Steel (2007)
  - Stefankovic and Vigoda (2007)
  - Mossel and Vigoda (2005)

Algebraic Methods for Proving Identifiability

Proposition

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two algebraic models. If there exist polynomials f_0 and f_1 such that

$$f_i(p) = 0 \textit{ for all } p \in \mathcal{M}_i, \textit{ and } f_i(p) \neq 0 \textit{ for some } p \in \mathcal{M}_{1-i}, \textit{ then}$$

$$\dim(\mathcal{M}_0 \cap \mathcal{M}_1) < \min(\dim \mathcal{M}_0, \dim \mathcal{M}_1).$$

Proposition

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two algebraic models. If there is a polynomial f_0 such that

$$f_0(p) = 0 \textit{ for all } p \in \mathcal{M}_0, \textit{ and } f_0(p) \neq 0 \textit{ for some } p \in \mathcal{M}_1, \textit{ and}$$

$$\dim \mathcal{M}_1 \leq \dim \mathcal{M}_0 \textit{ then}$$

$$\dim(\mathcal{M}_0 \cap \mathcal{M}_1) < \min(\dim \mathcal{M}_0, \dim \mathcal{M}_1).$$

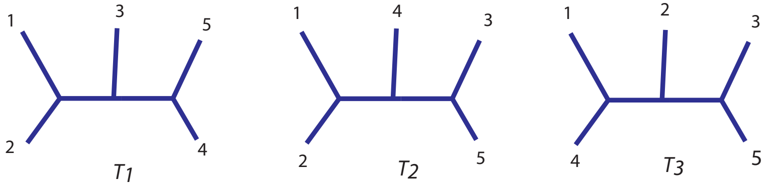
Our Identifiability Results

Theorem

The tree parameters of the phylogenetic 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 trivalent with $n \geq 4$ leaves.

- Strategy: Prove theorem for quartets $n = 4, 5, 6$.
- Use Matsen-Mossel-Steel “Six to Infinity" Theorem.
- Toric nature of group-based models lets us used tropical techniques to prove that models have the expected dimension.
- JC and K2P models allow us to construct linear invariants to prove identifiability.

A (Mathematical) Surprise



Theorem

For the Jukes-Cantor model

$$\overline{\mathcal{M}_{T_2}} \subseteq \overline{\mathcal{M}_{T_1} * \mathcal{M}_{T_3}}.$$

Can the closure be dropped; i.e. does it happen for biologically meaningful parameter values?

Future Directions

- Deal with the other group-based models (CFN, K3P)
- Beyond group-based models, GTR, GMM
- Beyond 2-tree mixtures to k-tree mixtures