Phylogenetic Algebraic Geometry

Seth Sullivant

North Carolina State University

March 15, 2012
Phylogenetics

Problem
Given a collection of species, find the tree that explains their history.

Data consists of aligned DNA sequences from homologous genes

Human: \[ \ldots \text{ACCGTGCAACGTGAACGA}\ldots \]
Chimp: \[ \ldots \text{ACCTTGGAAGGTAAACGA}\ldots \]
Gorilla: \[ \ldots \text{ACCGTGCAACGTAAACTA}\ldots \]
Model-Based Phylogenetics

- Use a probabilistic model of mutations
- Parameters for the model are the combinatorial tree $T$, and rate parameters for mutations on each edge
- Models give a probability for observing a particular aligned collection of DNA sequences

```plaintext
Human: ACCGTGCAACGTGAACGA
Chimp: ACGTTGCAAGGTAAACGA
Gorilla: ACCGTGCAACGTAAACTA
```

- Assuming site independence, data is summarized by empirical distribution of columns in the alignment.
  
  e.g. $\hat{p}(AAA) = \frac{6}{18}$, $\hat{p}(CGC) = \frac{2}{18}$, etc.

- Use empirical distribution and test statistic to find tree best explaining data
Phylogenetic Models

- Assuming site independence:
  - Phylogenetic Model is a latent class graphical model
  - Vertex $v \in T$ gives a random variable $X_v \in \{A, C, G, T\}$
  - All random variables corresponding to internal nodes are latent

$$P(x_1, x_2, x_3) = \sum_{y_1} \sum_{y_2} P(y_1) P(y_2 | y_1) P(x_1 | y_1) P(x_2 | y_2) P(x_3 | y_2)$$
Phylogenetic Models

- Assuming site independence:
- Phylogenetic Model is a latent class graphical model.
- Vertex $v \in T$ gives a random variable $X_v \in \{A, C, G, T\}$
- All random variables corresponding to internal nodes are latent.

\[
p_{i_1 i_2 i_3} = \sum_{j_1} \sum_{j_2} \pi_{j_1} a_{j_2,j_1} b_{i_1,j_1} c_{i_2,j_2} d_{i_3,j_2}
\]
Once we fix a tree $T$ and model structure, we get a map $\phi^T : \Theta \to \mathbb{R}^{4^n}$.

$\Theta \subseteq \mathbb{R}^d$ is a parameter space of numerical parameters (transition matrices associated to each edge).

The map $\phi^T$ is given by polynomial functions of the parameters.

For each $i_1 \cdots i_n \in \{A, C, G, T\}^n$, $\phi^T_{i_1 \cdots i_n}(\theta)$ gives the probability of the column $(i_1, \ldots, i_n)'$ in the alignment for the particular parameter choice $\theta$.

$$\phi^T_{i_1 i_2 i_3}(\pi, a, b, c, d) = \sum_{j_1} \sum_{j_2} \pi_{j_1} a_{j_2,j_1} b_{i_1,j_1} c_{i_2,j_2} d_{i_3,j_2}$$

The phylogenetic model is the set $\mathcal{M}_T = \phi^T(\Theta) \subseteq \mathbb{R}^{4^n}$. 
Let $\mathbb{R}[p] := \mathbb{R}[p_{i_1 \cdots i_n} : i_1 \cdots i_n \in \{A, C, G, T\}^n]$

**Definition**

Let

$$I_T := \langle f \in \mathbb{R}[p] : f(p) = 0 \text{ for all } p \in \mathcal{M}_T \rangle \subseteq \mathbb{R}[p].$$

$I_T$ is the **ideal of phylogenetic invariants** of $T$.

Let

$$V_T := \{p \in \mathbb{R}^{4n} : f(p) = 0 \text{ for all } f \in I_T\}.$$

$V_T$ is the **phylogenetic variety** of $T$.

- Note that $\mathcal{M}_T \subset V_T$.
- Since $\mathcal{M}_T$ is image of a polynomial map $\dim \mathcal{M}_T = \dim V_T$. 
\[ p_{lmno} = \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \pi i a_{ij} b_{ik} c_{jl} d_{jm} e_{kn} f_{ko} \]
\[ p_{lmno} = \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \pi_i a_{ij} b_{ik} c_{jl} d_{jm} e_{kn} f_{ko} \]

\[
\left( \sum_{j=1}^{4} a_{ij} c_{jl} d_{jm} \right)
\]
\[ p_{lmno} = \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \pi_i a_{ij} b_{ik} c_{jk} d_{jm} e_{kn} f_{ko} \]

\[
\left( \sum_{j=1}^{4} a_{ij} c_{jk} d_{jm} \right) \cdot \left( \sum_{k=1}^{4} b_{ik} e_{kn} f_{ko} \right)
\]
\[ p_{lmno} = \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \pi_i a_{ij} b_{ik} c_{jl} d_{jm} e_{kn} f_{ko} \]

\[ = \sum_{i=1}^{4} \pi_i \left( \left( \sum_{j=1}^{4} a_{ij} c_{jl} d_{jm} \right) \cdot \left( \sum_{k=1}^{4} b_{ik} e_{kn} f_{ko} \right) \right) \]
\[
p_{lmno} = \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \pi_i a_{ij} b_{ik} c_{jl} d_{jm} e_{kn} f_{ko}
\]

\[
= \sum_{i=1}^{4} \pi_i \left( \left( \sum_{j=1}^{4} a_{ij} c_{jl} d_{jm} \right) \cdot \left( \sum_{k=1}^{4} b_{ik} e_{kn} f_{ko} \right) \right)
\]

\[
\Rightarrow \text{rank} \left( \begin{array}{cccc}
p_{1111} & p_{1112} & \cdots & p_{1144} \\
p_{1211} & p_{1212} & \cdots & p_{1244} \\
\vdots & \vdots & \ddots & \vdots \\
p_{4411} & p_{4412} & \cdots & p_{4444}
\end{array} \right) \leq 4
\]
Definition

A **split** of a set is a bipartition \( A \mid B \). A split \( A \mid B \) of the leaves of a tree \( T \) is **valid** for \( T \) if the induced trees \( T\mid_A \) and \( T\mid_B \) do not intersect.

Example:

- **Valid**: \( 12 \mid 34 \)
- **Not Valid**: \( 13 \mid 24 \)
2-way Flattenings and Matrix Ranks

\[ p_{ijkl} = P(X_1 = i, X_2 = j, X_3 = k, X_4 = l) \]

\[
\text{Flat}_{12|34}(P) = \begin{pmatrix}
 p_{AAAA} & p_{AAAC} & p_{AAAG} & \cdots & p_{AATT} \\
 p_{ACAA} & p_{ACAC} & p_{ACAG} & \cdots & p_{ACTT} \\
 \vdots & \vdots & \vdots & \ddots & \vdots \\
 p_{TTAA} & p_{TTAC} & p_{TTAG} & \cdots & p_{TTTT}
\end{pmatrix}
\]

**Proposition**

Let \( P \in \mathcal{M}_T \).

- If \( A|B \) is a valid split for \( T \), then \( \text{rank}(\text{Flat}_{A|B}(P)) \leq 4 \).
  
  Invariants in \( I_T \) are subdeterminants of \( \text{Flat}_{A|B}(P) \).

- If \( C|D \) is not a valid split for \( T \), then generically \( \text{rank}(\text{Flat}_{C|D}(P)) > 4 \).
Phylogenetic Algebraic Geometry is the study of the phylogenetic varieties and ideals $V_T$ and $I_T$.

- Using Phylogenetic Invariants to Reconstruct Trees
- Identifiability of Phylogenetic Models
- Interesting Math—Useful in Other Problems
Using Phylogenetic Invariants to Reconstruct Trees

**Definition**

A phylogenetic invariant \( f \in I_T \) is **phylogenetically informative** if there is some other tree \( T' \) such that \( f \not\in I_{T'} \).

- Idea of Cavender-Felsenstein (1987), Lake (1987): Evaluate phylogenetically informative phylogenetic invariants at empirical distribution \( \hat{p} \) to reconstruct phylogenetic trees

**Proposition**

*For each n-leaf trivalent tree \( T \), let \( F_T \subseteq I_T \) be a set of phylogenetic invariants such that, for each \( T' \neq T \), there is an \( f \in F_T \), such that \( f' \not\in I_{T'} \). Let \( f_T := \sum_{f \in F_T} |f| \).

Then for generic \( p \in \cup M_T \), \( f_T(p) = 0 \) if and only if \( p \in M_T \).*

HOWEVER... Huelsenbeck only used linear invariants.

Casanellas, Fernandez-Sanchez (2006) redid these simulations using a generating set of the phylogenetic ideal $I_T$. Phylogenetic invariants become comparable to other methods.

For the particular model studied in Casanellas, Fernandez-Sanchez (2006) for a tree with 4 leaves, the ideal $I_T$ has 8002 generators.

$$f_T := \sum_{f \in F_T} |f|$$

is a sum of 8002 terms.

Major work to overcome combinatorial explosion for larger trees.
Definition

A parametric statistical model is identifiable if it gives 1-to-1 map from parameters to probability distributions.

“Is it possible to infer the parameters of the model from data?”

Identifiability guarantees consistency of statistical methods (ML)

Two types of parameters to consider for phylogenetic models:

- Numerical parameters (transition matrices)
- Tree parameter (combinatorial type of tree)
The unrooted tree parameter $T$ in a phylogenetic model is identifiable if for all $p \in \mathcal{M}_T$ there does not exist another $T' \neq T$ such that $p \in \mathcal{M}_{T'}$. 
The tree parameter in a phylogenetic model is \textit{generically identifiable} if for all \( n \)-leaf trees with \( T \neq T' \),

\[
\dim(\mathcal{M}_T \cap \mathcal{M}_{T'}) < \min(\dim(\mathcal{M}_T), \dim(\mathcal{M}_{T'})).
\]
Proposition

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

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

then 

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

The unrooted tree parameter of phylogenetic models is generically identifiable.

Proof.

- Edge flattening invariants can detect which splits are implied by a specific distribution in $\mathcal{M}_T$.
- The splits in $T$ uniquely determine $T$. 
Basic phylogenetic model assume same parameters at every site
- This assumption is not accurate within a single gene
  - Some sites more important: unlikely to change
- Tree structure may vary across genes
- Leads to mixture models for different classes of sites
- $\mathcal{M}(T, r)$ denotes a same tree mixture model with underlying tree $T$ and $r$ classes of sites
Question

For fixed number of trees $k$, are the tree parameters $T_1, \ldots, T_k$, and rate parameters of each tree (generically) identified in phylogenetic mixture models?

- $k = 1$ (Ordinary phylogenetic models)
  Most models are identifiable on $\geq 2, 3, 4$ leaves. (Rogers, Chang, Steel, Hendy, Penny, Székely, Allman, Rhodes, Housworth, ...)

- $k > 1$, $T_1 = T_2 = \cdots = T_k$ but no restriction on number of trees
  Not identifiable (Matsen-Steel, Stefankovic-Vigoda)

- $k > 1$, $T_i$ arbitrary
  Not identifiable (Mossel-Vigoda)
Theorem (Rhodes-Sullivant 2011)

The unrooted tree and numerical parameters in a $r$-class, same tree phylogenetic mixture model on $n$-leaf trivalent trees are generically identifiable, if $r < 4^{\lceil n/4 \rceil}$.

Proof Ideas.

- Phylogenetic invariants from flattenings
- Tensor rank (Kruskal’s Theorem) [Allman-Matias-Rhodes 2009]
- Elementary tree combinatorics
- Solving tree and numerical parameter identifiability at the same time
Theorem (Sturmfels-S, Allman-Rhodes, Casanellas-S, Draisma-Kuttler)

Consider “nice” algebraic phylogenetic model. The problem of computing phylogenetic invariants for any tree $T$ can be reduced to the same problem for star trees $K_{1,k}$.

- The ideal $I_T$ generated by local contributions from each $K_{1,k}$, plus flattening invariants from edges.
- The varieties $V_{K_{1,k}}$ are interesting classical algebraic varieties:
  - toric varieties
  - secant varieties
  - $Sec^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$
Random variables in finite abelian group $G$.

Transitions probabilities satisfy $\text{Prob}(X = g | Y = h) = f(g + h)$.

This means that the formula for $\text{Prob}(X_1 = g_1, \ldots, X_n = g_n)$ is a convolution (over $G^n$).

Apply discrete Fourier transform to turn convolution into a product.

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

*In the Fourier coordinates, a group-based model is parametrized by monomial functions in terms of the Fourier parameters. In particular, the CFN model is a toric variety.*
Theorem (Sturmfels-S 2005)

For any tree $T$, the toric ideal $I_T$ for the CFN model is generated by degree 2 determinantal equations.

Fourier coordinates:

$$q_{lmno} = \sum_{r,s,t,u \in \{0,1\}} (-1)^{rl+sm+tn+uo} \rho_{rstu}$$

$I_T$ generated by $2 \times 2$ minors of:

$$\begin{pmatrix} q_{0000} & q_{0001} & q_{0010} & q_{0011} \\ q_{1100} & q_{1101} & q_{1110} & q_{1111} \end{pmatrix} \begin{pmatrix} q_{0000} & q_{0011} \\ q_{0100} & q_{0111} \\ q_{1000} & q_{1011} \\ q_{1100} & q_{1111} \end{pmatrix} \begin{pmatrix} q_{0001} & q_{0010} \\ q_{0101} & q_{0110} \\ q_{1001} & q_{1010} \\ q_{1101} & q_{1110} \end{pmatrix}$$
Let $T = T_1 \# T_2$, tree obtained by joining two trees at a leaf.

Each ring $\mathbb{C}[p]/I_{T_1}$, $\mathbb{C}[p]/I_{T_2}$ is invariant under action of group $\mathcal{G} = \text{Gl}_r(\mathbb{C})^k$ acting on the glue leaves.

**Theorem (Draisma-Kuttler)**

- $\mathbb{C}[p]/I_T \cong (\mathbb{C}[p]/I_{T_1} \otimes_{\mathbb{C}} \mathbb{C}[p]/I_{T_2})^\mathcal{G}$
- $V_T = (V_{T_1} \times V_{T_2})//\mathcal{G}$ (GIT quotient)

Actions of individual factors ($\text{Gl}_r(\mathbb{C})$) do no interact.

Use Reynolds operator and first fundamental theorem of CIT.
Gluing more complex graphs

\[ \square + \square = \square \]

- Still a group action \((\text{Gl}_r(\mathbb{C})^k)\).
- But factors are not acting independently.
- \(\mathbb{C}[p]/I_G \ncong (\mathbb{C}[p]/I_{G_1} \otimes_{\mathbb{C}} \mathbb{C}[p]/I_{G_2})^G\)
- \(\mathbb{C}[p]/I_G \) generated by degree 1 part of \((\mathbb{C}[p]/I_{G_1} \otimes_{\mathbb{C}} \mathbb{C}[p]/I_{G_2})^G\)
  (toric fiber product if \(r = 1\))

Theorem (Engström-Kahle-S)

*Can determine generators of \(I_G\) from \(I_{G_1}\) and \(I_{G_2}\) if the TFP has “low codimension”.*

- Useful for other problems in algebraic statistics.
Phylogenetic models are fundamentally algebraic-geometric objects.

Algebraic perspective is useful for:
- Developing new construction algorithms
- Proving theorems about identifiability (currently best available for mixture models)

Leads to interesting new mathematics, useful for other problems

Long way to go: Your Help Needed!


A. Engström, T. Kahle, S. Sullivant. Multigraded commutative algebra of graph decompositions. 1102.2601


