Tying up loose strands: Defining equations of the strand symmetric model

Colby Long and Seth Sullivant

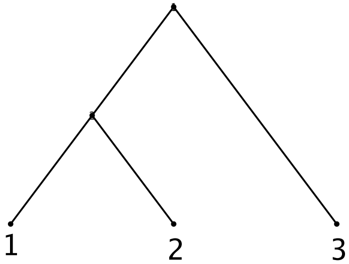
North Carolina State University

June 8, 2015

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
- Model evolution at a single locus.
- Give probability distribution on *n*-tuples of DNA characters

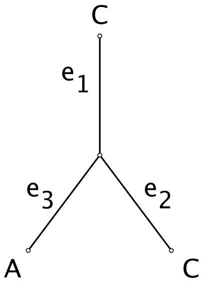
Phylogenetic Models

- Tree parameter: Binary leaf-labelled tree $\mathcal{T}$ with label set $[n]$.
- Random variable X_v associated to each node of $\mathcal{T}$.
- State space of each X_v is $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.
- Transition matrix associated to each edge.

$$M_{ij}^{k} = P(X_v = i | X_w = j).$$

- Entries of the transition matrices are the *stochastic* or *numerical parameters*.
- To find the probability of observing a particular state at the leaves, sum over all *histories*, the possible states of internal nodes.

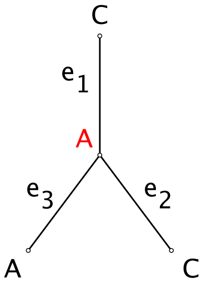
Jukes-Cantor Example



$$M^k = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_k & \beta_k & \beta_k & \beta_k \\ \beta_k & \alpha_k & \beta_k & \beta_k \\ \beta_k & \beta_k & \alpha_k & \beta_k \\ \beta_k & \beta_k & \beta_k & \alpha_k \end{pmatrix} \end{array}$$

$$M_{ij}^k = P(X_v = i | X_w = j)$$

Jukes-Cantor Example

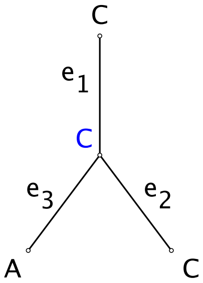


$$M^k = \begin{array}{c c} & \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{array}{c} A \\ C \\ G \\ T \end{array} & \begin{pmatrix} \alpha_k & \beta_k & \beta_k & \beta_k \\ \beta_k & \alpha_k & \beta_k & \beta_k \\ \beta_k & \beta_k & \alpha_k & \beta_k \\ \beta_k & \beta_k & \beta_k & \alpha_k \end{pmatrix} \end{array}$$

$$M^k_{ij} = P(X_v = i | X_w = j)$$

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

Jukes-Cantor Example

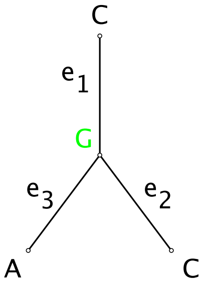


$$M^k = \begin{array}{c|cccc} & A & C & G & T \\ A & \alpha_k & \beta_k & \beta_k & \beta_k \\ C & \beta_k & \alpha_k & \beta_k & \beta_k \\ G & \beta_k & \beta_k & \alpha_k & \beta_k \\ T & \beta_k & \beta_k & \beta_k & \alpha_k \end{array}$$

$$M^k_{ij} = P(X_v = i | X_w = j)$$

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

Jukes-Cantor Example

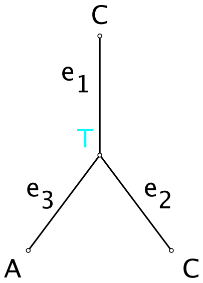


$$M^k = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & \alpha_k & \beta_k & \beta_k & \beta_k \\ C & \beta_k & \alpha_k & \beta_k & \beta_k \\ G & \beta_k & \beta_k & \alpha_k & \beta_k \\ T & \beta_k & \beta_k & \beta_k & \alpha_k \end{array}$$

$$M^k_{ij} = P(X_v = i | X_w = j)$$

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

Jukes-Cantor Example

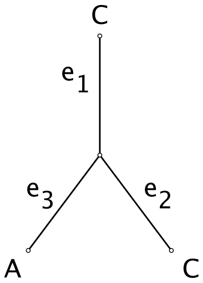


$$M^k = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & \alpha_k & \beta_k & \beta_k & \beta_k \\ C & \beta_k & \alpha_k & \beta_k & \beta_k \\ G & \beta_k & \beta_k & \alpha_k & \beta_k \\ T & \beta_k & \beta_k & \beta_k & \alpha_k \end{array}$$

$$M_{ij}^k = P(X_v = i | X_w = j)$$

$$p_{CCA} = \pi_A \beta_1 \beta_2 \alpha_3 + \pi_C \alpha_1 \alpha_2 \beta_3 + \pi_G \beta_1 \beta_2 \beta_3 + \pi_T \beta_1 \beta_2 \beta_3$$

Jukes-Cantor Example



$$M^k = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \alpha_k & \beta_k & \beta_k & \beta_k \\ \beta_k & \alpha_k & \beta_k & \beta_k \\ \beta_k & \beta_k & \alpha_k & \beta_k \\ \beta_k & \beta_k & \beta_k & \alpha_k \end{pmatrix} \end{array}$$

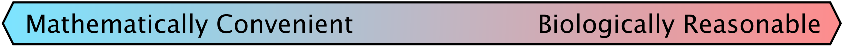
$$M^k_{ij} = P(X_v = i | X_w = j)$$

$$p_{CCA} = \pi_A \beta_1 \beta_2 \alpha_3 + \pi_C \alpha_1 \alpha_2 \beta_3 + \pi_G \beta_1 \beta_2 \beta_3 + \pi_T \beta_1 \beta_2 \beta_3$$

- $\psi_{\mathcal{T}} : \Theta_{\mathcal{T}} \to \Delta^{4^n - 1} \subseteq \mathbb{R}^{4^n}$
- $\mathcal{M}_{\mathcal{T}} = \psi_{\mathcal{T}}(\Theta_{\mathcal{T}})$ is the model.
- $V_{\mathcal{T}} = \overline{\mathrm{im}(\psi_{\mathcal{T}})}$ and $\mathcal{I}_{\mathcal{T}} = \mathcal{I}(V_{\mathcal{T}})$ is the ideal of phylogenetic invariants.

The Strand Symmetric Model (SSM)

- The Strand Symmetric Model (SSM) reflects the double-stranded structure of DNA.



Mathematically Convenient

Biologically Reasonable

- `A-T` and `C-G` are always paired, so a mutation in one induces a mutation in the other.
- We insist the root distribution satisfies $\pi_A = \pi_T$ and $\pi_C = \pi_G$.
- Likewise, if we let θ_{ij} be the entries of the transition matrices,

$$\begin{array}{cccc} \theta_{AA} = \theta_{TT} & \theta_{AC} = \theta_{TG} & \theta_{AG} = \theta_{TC} & \theta_{AT} = \theta_{TA} \\ \theta_{CC} = \theta_{GG} & \theta_{CG} = \theta_{GC} & \theta_{CT} = \theta_{GA} & \theta_{GT} = \theta_{CA} \end{array}$$

- Given any tree $\mathcal{T}$, we want to be able to determine $\mathcal{I}_{\mathcal{T}}$ for the SSM.

Determining the ideal of the SSM

Theorem (Casanellas-Sullivant 2005)

For any binary phylogenetic tree $\mathcal{T}$, the ideal of phylogenetic invariants for the SSM on $\mathcal{T}$ can be computed from the ideal of phylogenetic invariants for the claw tree, $\mathcal{I}_{SSM}$.

- Theoretically, this can be computed with elimination.
- Computing the required Gröbner basis is not possible.
- The Fourier transform gives a monomial parameterization for group-based models.
- We require something analogous for the Strand Symmetric Model.

Matrix-Valued Group-Based Models ([1])

- Identify states with elements of $\mathbb{Z}_2 \times \{0, 1\}$.
- $A = \binom{0}{0}$, $G = \binom{0}{1}$, $T = \binom{1}{0}$, $C = \binom{1}{1}$.

$$E = \begin{array}{ccc} & & \begin{array}{cccc} \multicolumn{2}{c}{0} & \multicolumn{2}{c}{1} \\ A & G & T & C \end{array} \\ \begin{array}{c} 0 \\ \\ 1 \\ \\ \end{array} & \begin{array}{c} A \\ G \\ T \\ C \end{array} & \left(\begin{array}{cc|cc} \theta_1 & \theta_8 & \theta_3 & \theta_2 \\ \theta_7 & \theta_5 & \theta_4 & \theta_6 \\ \hline \theta_3 & \theta_2 & \theta_1 & \theta_8 \\ \theta_4 & \theta_6 & \theta_7 & \theta_5 \end{array}\right) \end{array}$$

- $E_{i_1 i_2}^{j_1 j_2} = E_{i_1 i_2}^{k_1 k_2}$ whenever $j_1 - j_2 = k_1 - k_2$ in $\mathbb{Z}_2$.
- This makes the strand symmetric model a *matrix-valued group based model*.

The Group-Valued Fourier Transform

In the new coordinates, the parameterization of the cone over the SSM for $K_{1,3}$ is given by

$$q_{ijk}^{mno} = d_{0i}^{mm} e_{0j}^{nn} f_{0k}^{oo} + d_{1i}^{mm} e_{1j}^{nn} f_{1k}^{oo}$$

if $m + n + o \equiv 0$ in $\mathbb{Z}_2$, and $q_{ijk}^{mno} = 0$ otherwise.

- This is a projection of the space of rank 2 tensors.

$$Q = \begin{pmatrix} d_{00}^0 \\ d_{01}^0 \\ d_{00}^1 \\ d_{01}^1 \end{pmatrix} \otimes \begin{pmatrix} e_{00}^0 \\ e_{01}^0 \\ e_{00}^1 \\ e_{01}^1 \end{pmatrix} \otimes \begin{pmatrix} f_{00}^0 \\ f_{01}^0 \\ f_{00}^1 \\ f_{01}^1 \end{pmatrix} + \begin{pmatrix} d_{10}^0 \\ d_{11}^0 \\ d_{10}^1 \\ d_{11}^1 \end{pmatrix} \otimes \begin{pmatrix} e_{10}^0 \\ e_{11}^0 \\ e_{10}^1 \\ e_{11}^1 \end{pmatrix} \otimes \begin{pmatrix} f_{10}^0 \\ f_{11}^0 \\ f_{10}^1 \\ f_{11}^1 \end{pmatrix}$$

$$\mathcal{I}_{SSM} = \mathcal{I}(Sec^2(Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3))) \cap \mathbb{C}[q_{ijk}^{mno} : m + n + o = 0].$$

A Candidate Ideal

Using elimination, the same authors found $\mathcal{I}_{SSM}$ is generated by

- 32 equations in degree 3
- 18 equations in degree 4
- 0 equations in degree 5.
- Unknown for degree ≥ 6.

Theorem (L-Sullivant 2014)

Let $\mathcal{I}_F$ be the ideal generated by the 50 equations found in [1]. Then $\mathcal{I}_F = \mathcal{I}_{SSM}$.

We know that $\mathcal{I}_F \subseteq \mathcal{I}_{SSM}$ and $\mathcal{I}_{SSM}$ is prime, so we just need to show

1. $\dim(\mathcal{I}_F) = \dim(\mathcal{I}_{SSM})$.
2. $\mathcal{I}_F$ is prime.

How to show $\mathcal{I}_F$ is prime?

Dimension is easy,

- Compute $\dim(\mathcal{I}_F)$ with Macaulay2.
- Compute $\dim(\mathcal{I}_{SSM})$ as a tropical secant variety [3].
- $\dim(\mathcal{I}_F) = \dim(\mathcal{I}_{SSM}) = 20$.

How to show $\mathcal{I}_F$ is prime?

Dimension is easy,

- Compute $\dim(\mathcal{I}_F)$ with Macaulay2.
- Compute $\dim(\mathcal{I}_{SSM})$ as a tropical secant variety [3].
- $\dim(\mathcal{I}_F) = \dim(\mathcal{I}_{SSM}) = 20$.

Lemma [6, Proposition 23]

Let k be a field and $J \subset k[x_1, \ldots, x_n]$ be an ideal containing a polynomial $f = gx_1 + h$ with g, h not involving x_1 and g a non-zero divisor modulo J. Let $J_1 = J \cap k[x_2, \ldots, x_n]$ be the elimination ideal. Then J is prime if and only if J_1 is prime.

- J not prime $\Rightarrow$ J_1 not prime.
- Given $a, b \notin J$ with $ab \in J$, $a' := (ga - h_d x_1^{d-1} f) \notin J$, and $a'b \in J$ with lower x_1-degree.

Proving $\mathcal{I}_F$ is prime.

1. Start with $\mathcal{I}_0 = \mathcal{I}_F$ and $k = 1$.
2. Find a polynomial $f_k = g_k x_k + h_k \in \mathcal{I}_{k-1}$.
3. Verify that g_k is not a zero-divisor mod $\mathcal{I}_{k-1}$.
4. eliminate x_k to obtain the ideal $\mathcal{I}_k$.
5. Generate a decreasing chain of elimination ideals

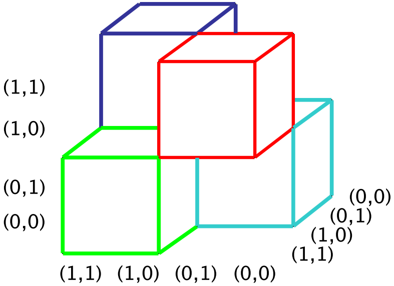
$$\mathcal{I}_F = \mathcal{I}_0 \supset \mathcal{I}_1 \supset \mathcal{I}_2 \ldots \supset \langle 0 \rangle.$$

By repeated application of the lemma,

$$\langle 0 \rangle \text{ prime } \Rightarrow \mathcal{I}_F \text{ prime }.$$

The Result

$$\mathcal{I}_{SSM} = \mathcal{I}(Sec^2(Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)))) \cap \mathbb{C}[q_{ijk}^{mno} : m + n + o = 0].$$



To reduce computation time...

- Take advantage of the group action on $\mathcal{I}_F$.
- Eliminate variables in particular order.

We show $\mathcal{I}_F = \mathcal{I}_{SSM}$ and therefore we can determine the ideal for the strand symmetric model for any binary tree $\mathcal{T}$.

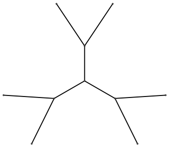
Another Application: CFN mixture models

- The CFN model is a two-state group-based phylogenetic model.
- Mixture models correspond to join varieties.

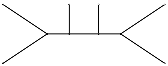
Goal

Find a generating set for the ideal of phylogenetic invariants for two-tree CFN mixtures on the same tree.

Snowflake



Caterpillar



- $\mathcal{I}_S * \mathcal{I}_S$ is generated by 32 equations in degree 3 and 18 equations in degree 4.

CFN Mixtures

- Relabeling coordinates, $\mathcal{I}_S * \mathcal{I}_S = \mathcal{I}_{SSM}$.
- We can also determine $\mathcal{I}_C * \mathcal{I}_C$

  1. Compute $\mathcal{I}_C * \mathcal{I}_C$ in degree 3 and 4.
  2. Apply Draisma tropical secant dimension [3].
  3. Apply the prime algorithm [6].

Observation

$HS(\mathcal{I}_C * \mathcal{I}_C, t) = HS(\mathcal{I}_S * \mathcal{I}_S, t)$.

Conjecture

For $\mathcal{T}, \mathcal{T}' \in \mathcal{T}_{[n]}$, $HS(\mathcal{I}_\mathcal{T} * \mathcal{I}_\mathcal{T}, t) = HS(\mathcal{I}_{\mathcal{T}'} * \mathcal{I}_{\mathcal{T}'}, t)$.

References

Marta Casanellas and Seth Sullivant.
Algebraic Statistics for Computational Biology, chapter 16.
Cambridge University Press, Cambridge, United Kingdom, 2005.

J.A. Cavender and J. Felsenstein.
Invariants of phylogenies in a simple case with discrete states.
J. of Class., 4:57–71, 1987.

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

Jan Draisma and Jochen Kuttler.
On the ideals of equivariant tree models.
Math. Ann., 344(3):619–644, 2009

S.N. Evans and T.P. Speed.
Invariants of some probability models used in phylogenetic inference.
Ann. Statist, 21(1):355–377, 1993.

Luis David Garcia, Michael Stillman, and Bernd Sturmfels.
Algebraic geometry of bayesian networks.
Journal of Symbolic Computation, 39(3-4):331–355, March-April 2005

D.R. Grayson and M.E. Stillman.
Macaulay2, a software system for research in algebraic geoemetry.
Available at http://www.math.uiuc.edu/Macaulay2/, 2002.

Colby Long and Seth Sullivant.
Tying up loose strands: Defining equations of the strand symmetric model.
Journal of Algebraic Statistics, 2015.