Small Phylogenetic Trees

M. Casanellas, M. Contois, L. D. Garcia, S. Hosten, Y. Kim, D. Levy, S. Snir

`lgpuente@msri.org`

MSRI

Objects

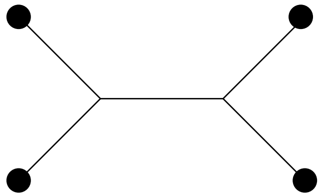
- Phylogenetic Trees with three, four, and five leaves.
- Rooted or un–rooted trees, with or without molecular clock assumption,
- Group models of evolution:
- Binary Symmetric $\begin{pmatrix} a_0 & a_1 \\ a_1 & a_0 \end{pmatrix}$, Jukes–Cantor $\begin{pmatrix} b & a & a & a \\ & b & a & a \\ & & b & a \\ & & & b \end{pmatrix}$,
- Kimura 2 $\begin{pmatrix} * & a & b & a \\ & * & a & b \\ & & * & a \\ & & & * \end{pmatrix}$, Kimura 3 $\begin{pmatrix} * & a & b & c \\ & * & c & b \\ & & * & a \\ & & & * \end{pmatrix}$.

Goals

- Describe the model parameterization
  - in the probability simplex,
  - in the Fourier coordinates.
- Compute
  - dimension – least number of parameters needed to describe the model,
  - degree,
  - embedding dimension – sufficient statistics,
  - singular locus (its dimension and degree),
  - ML degree,
  - MLE.
- Develop an alternative analytic method for tree reconstruction.
- Comparison between analytic method and numerical methods like DNAml.
- Create a web page to make technology available to computational biologists.

Parameterization in the probability simplex

- Kimura 2 model on the quartet un–rooted tree.

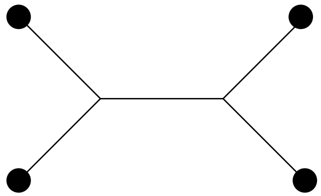


- Order the bases as A, G, C, T. Attached to each edge e, there is a symmetric matrix M_e equal to

$$\begin{pmatrix} c_e & a_e & b_e & b_e \\ & c_e & b_e & b_e \\ & & c_e & a_e \\ & & & c_e \end{pmatrix}$$

Parameterization in the probability simplex

- Kimura 2 model on the quartet un–rooted tree.



- The probability of observing i, j, k, l at the leaves equals

$$p_{ijkl} = \sum_{(w_1,w_2)\in\{A,G,C,T\}^2} M_1(w_1, i)M_2(w_1, j)M_3(w_2, k)M_4(w_2, l)M_5(w_1, w_2)$$

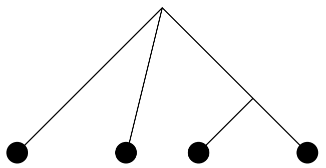
- For any $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ based model we have

$$p_{ijk} = p_{ijk1} = p_{(i+2)(j+2)(k+2)2} = p_{(i+3)(j+3)(k+3)3} = p_{(i+4)(j+4)(k+4)4}.$$

- For example $p_{CCC} = p_{CCCA} = p_{TTTG} = p_{AAAC} = p_{GGGT}$.
- Hence, the embedding dimension of the model is less or equal to 64.

Fourier parameterization

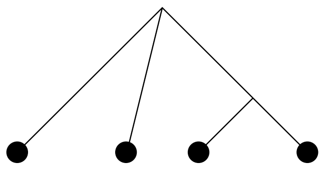
- Consider the "giraffe" model on four taxa with uniform root distribution and molecular clock.



- Note that without molecular clock, both models are equivalent.
- The Fourier transformation is a linear map that *simultaneously* diagonalizes all matrices M_e. So we have five diagonal 4×4–matrices X, Y, Z, V, W.
- The Fourier parameters are denoted q_{ijk} representing q_{ijkl}, where $l = i + j + k$.

Fourier parameterization

- Consider the “giraffe” model on four taxa with uniform root distribution and molecular clock.



- The Fourier parameterization is the **monomial** parameterization

$$q_{ijk} = x_i y_j z_{k+l} v_k w_l = x_i y_j z_{i+j} v_k w_{i+j+k}.$$

- The Kimura 2 assumption implies

$$x_3 = x_4, \quad y_3 = y_4, \quad z_3 = z_4, \quad v_3 = v_4, \quad w_3 = w_4.$$

- The molecular clock assumption implies $X = Y$, $V = W$, $X = ZW$, that is

$$x_i = y_i, \quad v_i = w_i, \quad x_i = v_i z_i.$$

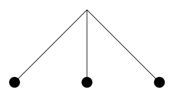
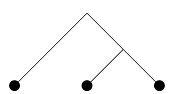
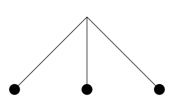
- The **binomial** ideal $I = \mathrm{toric{-}ideal}(\mathrm{monomial\ map})$ is the ideal of polynomial invariants in the Fourier parameters.

Solving the likelihood equations

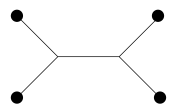
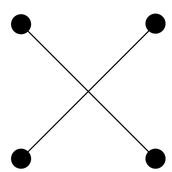
$$I \longrightarrow M_I \longrightarrow K = \ker(M_I) \longrightarrow I_{K,u} \longrightarrow J = \mathrm{sat}(I_{K,u}, \mathrm{slocus}(I)).$$

- Kernel of a polynomial matrix:
  - Linear algebra approach to compute kernel (HMM group).
  - Smaller matrices: Enough $\mathrm{codim}(I)$ equations to do computations.
  - Direct computations on the Fourier parameters.
  - Homotopy methods (PHC) to avoid kernel computation.
- Lower bounds for ML degree: Taking a subcollection of the rows of M_I.
- Upper bounds for ML degree:
  - Degree of zero-dimensional $I_{K,u}$ before saturation,
  - ML degree bounded by a sum of mixed volumes of Newton polytopes of the polynomial parameterization.

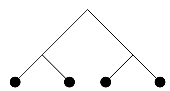
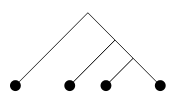
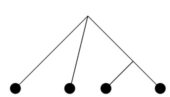
Trees with three leaves

		d	ed	m	sd	sm	MLd
	BS	4	7	8	1	24	92
	JC	3	4	3	1	3	23
	K2	6	9	12	3	22	
	K3	9	15	96			
	BS	2	2	1	-	-	1
	JC	2	3	13	1	1	15
	K2	4	6	6	2	10	190
	K3	6	9	12	3	22	
	BS	1	1	1	-	-	1
	JC	1	2	3	0	2	7
	K2	2	3	3	1	1	15
	K3	3	4	3	1	3	40

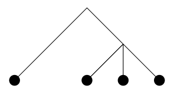
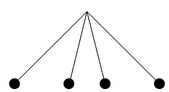
Trees with four leaves no molecular clock

		d	ed	m	sd	sm	MLd
	BS	5	7	4	2	4	14
	JC	5	14				
	K2	10					
	K3	15	63				
	BS	4	7	8	1	24	92
	JC	4					
	K2	8					
	K3	12					

Trees with four leaves molecular clock

		d	ed	m	sd	sm	MLd
	BS	3	4 (7)	2	1	1	1
	JC	3		14			
	K2	6		108			
	K3	9		1619			
	BS	3	4 (7)	2	1	1	9
	JC	3		14			
	K2	6		129			
	K3	9		1619			
	BS	2	7	2	0	1	6
	JC	2		11			
	K2	4		45			
	K3	6		227			

Trees with four leaves molecular clock

		d	ed	m	sd	sm	MLd
	BS	2	3	2	0	1	3
	JC	2		5			
	K2	4		18			
	K3	6		80			
	BS	1	2	2	0	1	3
	JC	1		4	0	2	
	K2	2		8			
	K3	3		16			