

Phylogenetic algebraic geometry

Nicholas Eriksson

Department of Mathematics
University of California, Berkeley

6 December 2005

Primera Escuela Argentina de Matemática y Biología

Outline

Phylogenetic
algebraic
geometry

Nicholas Eriksson

Markov models on trees

Markov models on
trees

Secant varieties of Segre varieties

Secant varieties of
Segre varieties

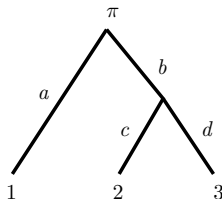
Group based models

Group based
models

Maximum likelihood

Maximum
likelihood

Trees



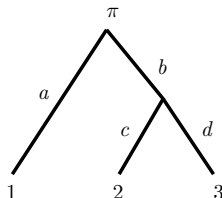
$$\pi = (\pi_0, \pi_1)$$

$$M_a = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \quad M_b = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix}$$

$$M_c = \begin{pmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{pmatrix} \quad M_d = \begin{pmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{pmatrix}$$

$$\pi_u \cdot a_{ui} \cdot b_{uv} \cdot c_{vj} \cdot d_{vk}$$

Parameterization



$$p_{ijk} = Pr(X_1 = i, X_2 = j, X_3 = k) =$$
$$\pi_0 a_{0i} b_{00} c_{0j} d_{0k} + \pi_0 a_{0i} b_{01} c_{1j} d_{1k} +$$
$$\pi_1 a_{1i} b_{10} c_{0j} d_{0k} + \pi_1 a_{1i} b_{11} c_{1j} d_{1k}$$

This parameterizes a variety in 8 dimensional space. Which variety is it?

We can solve such problems using Gröbner bases. For example take the variety given by 2 parameters a, b :

$$f_1 = a^4$$

$$f_2 = a^3b$$

$$f_3 = ab^3$$

$$f_4 = b^4$$

What are the equations that $f_1, f_2, f_3,$ and f_4 satisfy?

Gröbner bases

Solution: start with the ideal

$$\langle f_1 - a^4, f_2 - a^3b, f_3 - ab^3, f_4 - b^4 \rangle$$

and start “eliminating” variables

A Gröbner basis:

$$\langle f_2f_3 - f_1f_4, bf_3 - af_4, bf_1 - af_2, f_3^3 - f_2f_4^2, f_1f_3^2 - f_2^2f_4, \\ af_3^2 - bf_2f_4, f_2^3 - f_1^2f_3, bf_2^2 - af_1f_3, b^2f_2 - a^2f_3, \\ b^4 - f_4, ab^3 - f_3, a^3b - f_2, a^4 - f_1, a^2b^2f_4 - f_3^2 \rangle$$

This lets us eliminate a, b to get:

$$\langle f_2f_3 - f_1f_4, f_3^3 - f_2f_4^2, f_1f_3^2 - f_2^2f_4, f_2^3 - f_1^2f_3 \rangle$$

Restricting the model

Assume that the four transition matrices are identical

$$M_a = M_b = M_c = M_d = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}.$$

This model satisfies linear equations

$$p_{001} = p_{010} \quad \text{and} \quad p_{101} = p_{110}$$

and a degree eight polynomial with 70 terms.

Other models

- General** no constraints on the k^2 entries of M_e .
- Group based** The matrices M_e are simultaneously diagonalizable by the Fourier transform of an abelian *group*.
- Stationary** The matrices M_e all share the common left eigenvector $\pi = (\pi_1, \dots, \pi_k)$.
- Reversible** The matrices M_e are *symmetric* with the common left eigenvector $\pi = (1, 1, \dots, 1)$.
- Commuting** The matrices M_e commute pairwise.
- Substitution** The M_e matrices have the form $\exp(t_e \cdot Q)$ where Q is a fixed matrix. This is the most widely used model in biology

More models

Homogeneous The matrices M_e are all equal, or they all belong to a small finite collection.

No hidden nodes When all nodes are observed random variables then the parameterization becomes monomial, and the model is a toric variety.

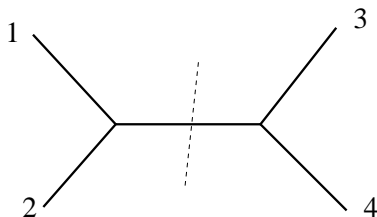
Mixtures The sum of these several different models. For example, this may be used to model the fact that different regions of the genome evolve at different rates.

The problems

- ▶ What are the interesting phylogenetic models?
- ▶ What are the ideals of these models?
- ▶ Study the geometry of these ideals (degree, dimension, singular locus, ...)

Theorem (Allman-Rhodes). The ideal of the general Markov model with binary random variables on a binary tree T is generated by the 3×3 subdeterminants of all “flattenings along edges of T ” of the table $p_{i_1 \dots i_n}$

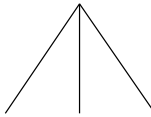
Flattening



$$\begin{array}{c} 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{pmatrix} p_{0000} & p_{0001} & p_{0010} & p_{0011} \\ p_{0100} & p_{0101} & p_{0110} & p_{0111} \\ p_{1000} & p_{1001} & p_{1010} & p_{1011} \\ p_{1100} & p_{1101} & p_{1110} & p_{1111} \end{pmatrix}.$$

The crucial case

Theorem (Allman and Rhodes). Knowledge of the ideal

of the tree  is the key problem in finding the
ideals of bigger trees (under the general Markov model).

We saw that for binary random variables, this ideal is
empty.

The next challenge: $\{A, C, G, T\}$. This is a **secant variety**.

Secant varieties

Phylogenetic
algebraic
geometry

Nicholas Eriksson

Markov models on
trees

Secant varieties of
Segre varieties

Group based
models

Maximum
likelihood

The secant variety of a variety X is the set of all points that lie on a line passing through two points of X .

The secant variety of a variety X is the set of all points that lie on a line passing through two points of X .

The key to understanding DNA evolution under the general Markov model is the ideal of the fourth secant variety of a Segre variety:

$$X = \text{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$$

A group action

- ▶ GL_n is the set of invertible $n \times n$ matrices
- ▶ GL_4 acts on \mathbb{R}^4 by multiplication.
- ▶ This means that $GL_4 \times GL_4 \times GL_4$ acts on X .
- ▶ So we can break X up into irreducible parts.

Representations of GL_n

- ▶ A partition of 5: $3 + 1 + 1$

Representations of GL_n

- ▶ A partition of 5: $3 + 1 + 1$
- ▶ Irreducible representations of GL_n are indexed by partitions of n .

Representations of GL_n

- ▶ A partition of 5: $3 + 1 + 1$
- ▶ Irreducible representations of GL_n are indexed by partitions of n .
- ▶ Write $S_\lambda \mathbb{C}^n$ for the irreducible representation of GL_n indexed by the partition λ .

$$\dim S_\lambda \mathbb{C}^n = \prod \frac{n - i + j}{h_{ij}},$$

where the product is over all boxes in the Young diagram of λ , and h_{ij} is the *hook length* of λ at i, j

Representations of GL_n

- ▶ A partition of 5: $3 + 1 + 1$
- ▶ Irreducible representations of GL_n are indexed by partitions of n .
- ▶ Write $S_\lambda \mathbb{C}^n$ for the irreducible representation of GL_n indexed by the partition λ .

$$\dim S_\lambda \mathbb{C}^n = \prod \frac{n - i + j}{h_{ij}},$$

where the product is over all boxes in the Young diagram of λ , and h_{ij} is the *hook length* of λ at (i, j)

Hook lengths	Coordinates (i,j)
5 2 1	1,1 1,2 2,2
2	2,1
1	3,1

- ▶ Example:

$$\dim S_{311} \mathbb{C}^n = \frac{(n+2)(n+1)n(n-1)(n-2)}{20},$$

so $\dim S_{311} \mathbb{C}^4 = 36$.

Degree 5 and 9

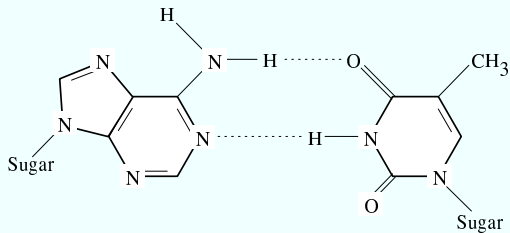
The ideal of $X = \text{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$ contains a 1728 dimensional space of degree 5 polynomials

$$\begin{aligned} & S_{311}A \otimes S_{2111}B \otimes S_{2111}C \\ & \oplus S_{2111}A \otimes S_{311}B \otimes S_{2111}C \\ & \oplus S_{2111}A \otimes S_{2111}B \otimes S_{311}C, \end{aligned}$$

It also contains $S_{333}A \otimes S_{333}B \otimes S_{333}C$ in degree 9, which is not generated by the degree 5 polynomials.

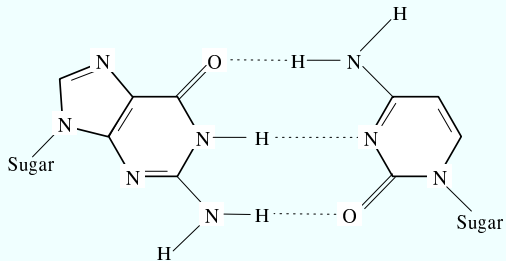
Open problem. Find any other information about the ideal of X .

DNA review



Adenine (A)

Thymine (T)



Guanine (G)

Cytosine (C)

The models

Jukes-Cantor binary $\begin{pmatrix} \cdot & a \\ a & \cdot \end{pmatrix}$

Jukes-Cantor DNA $\begin{pmatrix} \cdot & a & a & a \\ a & \cdot & a & a \\ a & a & \cdot & a \\ a & a & a & \cdot \end{pmatrix}$

Kimura 2 parameter $\begin{pmatrix} \cdot & b & a & b \\ b & \cdot & b & a \\ a & b & \cdot & b \\ b & a & b & \cdot \end{pmatrix}$

Kimura 3 parameter $\begin{pmatrix} \cdot & b & a & c \\ b & \cdot & c & a \\ a & c & \cdot & b \\ c & a & b & \cdot \end{pmatrix}$

Why group based?

The multiplication table for the group

$\mathbb{Z}/(2) \times \mathbb{Z}/(2) = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ is

	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(0,0)	(0,1)	(1,0)	(1,1)
(0,1)	(0,1)	(0,0)	(1,1)	(1,0)
(1,0)	(1,0)	(1,1)	(0,0)	(0,1)
(1,1)	(1,1)	(1,0)	(0,1)	(0,0)

Why group based?

The multiplication table for the group

$\mathbb{Z}/(2) \times \mathbb{Z}/(2) = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ is

	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(0,0)	(0,1)	(1,0)	(1,1)
(0,1)	(0,1)	(0,0)	(1,1)	(1,0)
(1,0)	(1,0)	(1,1)	(0,0)	(0,1)
(1,1)	(1,1)	(1,0)	(0,1)	(0,0)

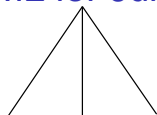
This looks like the Kimura 3 parameter matrix!

$$\begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}$$

For (much) more, see “Toric ideals of phylogenetic invariants”, Sturmfels & Sullivant,

<http://arxiv.org/q-bio.PE/0402015>

ML for Jukes-Cantor



$$M_i = \begin{pmatrix} 1 - 3\theta_1 & \theta_i & \theta_i & \theta_i \\ \theta_i & 1 - 3\theta_i & \theta_i & \theta_i \\ \theta_i & \theta_i & 1 - 3\theta_i & \theta_i \\ \theta_i & \theta_i & \theta_i & 1 - 3\theta_i \end{pmatrix}$$

for $i = 1, 2, 3$.

$$f_1(\theta) = -24\theta_1\theta_2\theta_3 + 9\theta_1\theta_2 + 9\theta_1\theta_3 + \\ 9\theta_2\theta_3 - 3\theta_1 - 3\theta_2 - 3\theta_3 + 1,$$

$$f_2(\theta) = -48\theta_1\theta_2\theta_3 + 6\theta_1\theta_2 + 6\theta_1\theta_3 + 6\theta_2\theta_3,$$

$$f_3(\theta) = 24\theta_1\theta_2\theta_3 + 3\theta_1\theta_2 - 9\theta_1\theta_3 - 9\theta_2\theta_3 + 3\theta_3,$$

$$f_4(\theta) = 24\theta_1\theta_2\theta_3 - 9\theta_1\theta_2 + 3\theta_1\theta_3 - 9\theta_2\theta_3 + 3\theta_2,$$

$$f_5(\theta) = 24\theta_1\theta_2\theta_3 - 9\theta_1\theta_2 - 9\theta_1\theta_3 + 3\theta_2\theta_3 + 3\theta_1.$$

Log-likelihood

The log-likelihood equation:

Given data $u = (u_1, \dots, u_5)$, the likelihood of observing u with parameters θ is

$$\ell_u(\theta) = \sum_{i=1}^5 u_i \cdot \log(f_i(\theta))$$

Log-likelihood

The log-likelihood equation:

Given data $u = (u_1, \dots, u_5)$, the likelihood of observing u with parameters θ is

$$\ell_u(\theta) = \sum_{i=1}^5 u_i \cdot \log(f_i(\theta))$$

Maximum likelihood

Given data u , find parameters θ that maximize $\ell_u(\theta)$.

Solve the critical equations:

$$\frac{\partial \ell_u}{\partial \theta_1} = \frac{\partial \ell_u}{\partial \theta_2} = \dots = \frac{\partial \ell_u}{\partial \theta_d} = 0.$$

Polynomials

$$\frac{\partial \ell_u}{\partial \theta_i} = \frac{u_1}{f_1(\theta)} \frac{\partial f_1}{\partial \theta_i} + \frac{u_2}{f_2(\theta)} \frac{\partial f_2}{\partial \theta_i} + \cdots + \frac{u_5}{f_5(\theta)} \frac{\partial f_5}{\partial \theta_i}.$$

is a rational function.

After clearing denominators, solving

$$\frac{\partial \ell_u}{\partial \theta_1} = \frac{\partial \ell_u}{\partial \theta_2} = \cdots = \frac{\partial \ell_u}{\partial \theta_d} = 0.$$

involves solving a system of polynomial equations.