Using algebraic geometry for phylogenetic reconstruction

Marta Casanellas i Rius
(joint work with Jesús Fernández-Sánchez)

Departament de Matemàtica Aplicada I
Universitat Politècnica de Catalunya

IMA Workshop
Applications in Biology, Dynamics, and Statistics
March 8, 2007
 Outline

1. Algebraic evolutionary models
2. Phylogenetic inference using algebraic geometry
3. The geometry of the Kimura variety
4. New results on simulated data

M. Casanellas (UPC)
Outline

1. Algebraic evolutionary models
2. Phylogenetic inference using algebraic geometry
3. The geometry of the Kimura variety
4. New results on simulated data
Outline

1. Algebraic evolutionary models
2. Phylogenetic inference using algebraic geometry
3. The geometry of the Kimura variety
4. New results on simulated data
Outline

1. Algebraic evolutionary models
2. Phylogenetic inference using algebraic geometry
3. The geometry of the Kimura variety
4. New results on simulated data
Outline

1. Algebraic evolutionary models
2. Phylogenetic inference using algebraic geometry
3. The geometry of the Kimura variety
4. New results on simulated data
Phylogenetic reconstruction

Given $n$ current species, e.g. HUMAN, GORILLA, CHIMP,

- given part of their genome and an alignment
Phylogenetic reconstruction

Given $n$ current species, e.g. HUMAN, GORILLA, CHIMP,
- given part of their genome and an alignment

Goal: to reconstruct their ancestral relationships (phylogeny):
Mutations, deletions and insertions of nucleotides occur along the speciation process.

Given two DNA sequences, an alignment is a correspondence between them that accounts for their differences. The optimal alignment is the one that minimizes the number of mutations, deletions and insertions.

\[ \text{seq}_1 : ACGTAGCTAAGTTA... \quad \text{seq}_2 : ACCGAGACCCAGTA... \]

A possible alignment is:

\[
\begin{align*}
\end{align*}
\]
Mutations, deletions and insertions of nucleotides occur along the speciation process.

Given two DNA sequences, an alignment is a correspondence between them that accounts for their differences. The optimal alignment is the one that minimizes the number of mutations, deletions and insertions.

\[ \text{seq}_1 : AC\underline{GTA}\underline{GCTA}A\ldots \quad \text{seq}_2 : ACCGAGACCCAGTA\ldots \]

A possible alignment is:

\[
\begin{align*}
\text{seq}_1 & \quad A & C & - & G & - & T & A & - & G & C & T & A & A & G & T & T & T & A \\
\text{seq}_2 & \quad A & C & C & G & A & G & A & C & - & C & C & A & - & G & T & - & A
\end{align*}
\]
Phylogenetic reconstruction

Given $n$ current species, e.g. HUMAN, GORILLA, CHIMP, given part of their genome and an alignment

```
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AAC</td>
<td>TTCAAGGCTTACCGCTG</td>
</tr>
<tr>
<td>AAGGTCGATGCTCACCAGATG</td>
<td></td>
</tr>
<tr>
<td>AACGTC</td>
<td>TATGCTCACCAGATG</td>
</tr>
</tbody>
</table>
```

Goal: to reconstruct their ancestral relationships (phylogeny):
Phylogenetic reconstruction

Given $n$ current species, e.g. HUMAN, GORILLA, CHIMP,
- given part of their genome and an alignment

$\begin{align*}
\text{AAC} & \text{TTCGAGGCTTACCGCTG} \\
\text{AAGGTCGATGCTCACCAGATG} \\
\text{AACGTC} & \text{TATGCTCACCAGATG}
\end{align*}$

Goal: to reconstruct their ancestral relationships (phylogeny):

\[ \text{Diagram of ancestral relationships} \]
Algebraic evolutionary models

- Assume that all sites of the alignment evolve equally and independently.
- At each node of the tree we put a random variable taking values in \{A, C, G, T\}
- Variables at the leaves are **observed** and variables at the interior nodes are **hidden**.
- At each branch we write a matrix (**substitution matrix**) with the probabilities of a nucleotide at the parent node being substituted by another at its child.

\[
S = \begin{pmatrix}
A & P(A|A) & P(C|A) & P(G|A) & P(T|A) \\
C & P(A|C) & P(C|C) & P(G|C) & P(T|C) \\
G & P(A|G) & P(C|G) & P(G|G) & P(T|G) \\
T & P(A|T) & P(C|T) & P(G|T) & P(T|T)
\end{pmatrix}
\]

- \(t_i = \text{branch length}\)
  (represents the number of mutations per site along that branch)
Algebraic evolutionary models

- Assume that all sites of the alignment evolve equally and independently.
- At each node of the tree we put a random variable taking values in \{A, C, G, T\}
  \[t_i = \text{branch length}\]
  (represents the number of mutations per site along that branch)
- Variables at the leaves are observed and variables at the interior nodes are hidden.
- At each branch we write a matrix (substitution matrix) with the probabilities of a nucleotide at the parent node being substituted by another at its child.

\[
S = \begin{pmatrix}
A & \mathbb{P}(A|A) & \mathbb{P}(C|A) & \mathbb{P}(G|A) & \mathbb{P}(T|A) \\
C & \mathbb{P}(A|C) & \mathbb{P}(C|C) & \mathbb{P}(G|C) & \mathbb{P}(T|C) \\
G & \mathbb{P}(A|G) & \mathbb{P}(C|G) & \mathbb{P}(G|G) & \mathbb{P}(T|G) \\
T & \mathbb{P}(A|T) & \mathbb{P}(C|T) & \mathbb{P}(G|T) & \mathbb{P}(T|T)
\end{pmatrix}
\]
The entries of $S_i$ are unknown parameters.

Example (Group-based models. Kimura 3-parameters)

- Root node has uniform distribution: $\pi_A = \ldots = \pi_T = 0.25$

\[
A \quad C \quad G \quad T
\]

\[
S_i = \begin{pmatrix}
    a_i & b_i & c_i & d_i \\
    b_i & a_i & d_i & c_i \\
    c_i & d_i & a_i & b_i \\
    d_i & c_i & b_i & a_i
\end{pmatrix}, \quad a_i + b_i + c_i + d_i = 1
\]

- Kimura 2-parameters: $b_i = d_i$.
- Jukes-Cantor: $b_i = c_i = d_i$. 

[M. Casanellas (UPC)]

Using algebraic geometry ...

IMA, March 2007 8 / 36
The entries of $S_i$ are unknown parameters.

**Example (Group-based models. Kimura 3-parameters)**

- Root node has uniform distribution: $\pi_A = \ldots = \pi_T = 0.25$
  
  $$
  \begin{array}{cccc}
  A & C & G & T \\
  a_i & b_i & c_i & d_i \\
  b_i & a_i & d_i & c_i \\
  c_i & d_i & a_i & b_i \\
  d_i & c_i & b_i & a_i \\
  \end{array}
  $$

  $$
  S_i = 
  \begin{pmatrix}
  a_i & b_i & c_i & d_i \\
  b_i & a_i & d_i & c_i \\
  c_i & d_i & a_i & b_i \\
  d_i & c_i & b_i & a_i \\
  \end{pmatrix}, \quad a_i + b_i + c_i + d_i = 1
  $$

- Kimura 2-parameters: $b_i = d_i$.

- Jukes-Cantor: $b_i = c_i = d_i$. 

---

**Diagram**: Shows transitions and transversions between purines (A, G) and pyrimidines (T, C).
• The entries of $S_i$ are unknown parameters.

Example (Group-based models. Kimura 3-parameters)

• Root node has uniform distribution: $\pi_A = ... = \pi_T = 0.25$

\[
S_i = \begin{pmatrix}
A & C & G & T \\
A & b_i & c_i & d_i \\
C & a_i & b_i & c_i \\
G & c_i & d_i & a_i \\
T & d_i & c_i & b_i \\
\end{pmatrix}, \quad a_i + b_i + c_i + d_i = 1
\]

• Kimura 2-parameters: $b_i = d_i$.

• Jukes-Cantor: $b_i = c_i = d_i$. 
Hidden Markov process

We denote the joint distribution of the observed variables $X_1, X_2, X_3$ as $p_{x_1 x_2 x_3} = \text{Prob}(X_1 = x_1, X_2 = x_2, X_3 = x_3)$.

\[ p_{x_1 x_2 x_3} = \sum_{y_4, y_r \in \{A, C, G, T\}} \pi_{y_r} S_1(x_1, y_r) S_4(y_4, y_r) S_2(x_2, y_4) S_3(x_3, y_4) \]

- $p_{x_1 x_2 x_3}$ is a homogeneous polynomial on the parameters whose degree is the number of edges.
The **evolutionary model** defines a polynomial map

$$
\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{4n}
$$

$$
\theta = (\theta_1, \ldots, \theta_d) \mapsto (p_{AA\ldots A}, p_{AA\ldots C}, p_{AA\ldots G}, \ldots, p_{TT\ldots T})
$$

$$
\varphi : \triangle^{d-1} \rightarrow \triangle^{4n-1}
$$

$$
\varphi : \mathbb{C}^d \rightarrow \mathbb{C}^{4n}
$$

**Algebraic variety** \( V = \overline{\text{im}\varphi} \), closure in Zariski topology.

- Given an alignment, we can estimate the joint probability \( p_{x_1 x_2 x_3} \) as the relative frequency of column \( x_1 x_2 x_3 \) in the alignment.
- In the theoretical model, this would be a point on the variety \( V \).
- Goal: use the ideal of \( V \), \( I(V) \subset \mathbb{R}[p_{AA\ldots A}, p_{AA\ldots C}, \ldots, p_{TT\ldots T}] \), to infer the topology of the phylogenetic tree.
The evolutionary model defines a polynomial map

\[ \varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{4n} \]

\[ \theta = (\theta_1, \ldots, \theta_d) \mapsto (p_{AA\ldots A}, p_{AA\ldots C}, p_{AA\ldots G}, \ldots, p_{TT\ldots T}) \]

\[ \varphi : \Delta^{d-1} \rightarrow \Delta^{4n-1} \]

\[ \varphi : \mathbb{C}^d \rightarrow \mathbb{C}^{4n} \]

Algebraic variety \( V = \text{im} \varphi \), closure in Zariski topology.

Given an alignment, we can estimate the joint probability \( p_{x_1,x_2,x_3} \) as the relative frequency of column \( x_1 x_2 x_3 \) in the alignment.

In the theoretical model, this would be a point on the variety \( V \).

Goal: use the ideal of \( V \), \( I(V) \subset \mathbb{R}[p_{AA\ldots A}, p_{AA\ldots C}, \ldots, p_{TT\ldots T}] \), to infer the topology of the phylogenetic tree.
The evolutionary model defines a polynomial map

\[ \varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{4n}, \]

\[ \theta = (\theta_1, \ldots, \theta_d) \mapsto (p_{\text{AA}A}, p_{\text{AA}C}, p_{\text{AA}G}, \ldots, p_{\text{TT}T}) \]

\[ \varphi : \triangle^{d-1} \rightarrow \triangle^{4n-1} \]

\[ \varphi : \mathbb{C}^d \rightarrow \mathbb{C}^{4n} \]

Algebraic variety \( V = \overline{\text{im} \varphi} \), closure in Zariski topology.

Given an alignment, we can estimate the joint probability \( p_{x_1,x_2,x_3} \) as the relative frequency of column \( x_1x_2x_3 \) in the alignment.

In the theoretical model, this would be a point on the variety \( V \).

Goal: use the ideal of \( V \), \( I(V) \subset \mathbb{R}[p_{\text{AA}A}, p_{\text{AA}C}, \ldots, p_{\text{TT}T}] \), to infer the topology of the phylogenetic tree.
Computing the ideal

(Eriksson–Ranestad–Sturmfels–Sullivant ’05)

- Some generators of $I(V)$ depend only on the model chosen (not on the topology). E.g. for Jukes-Cantor they are: $\sum p_{x_1x_2x_3} = 1$

$$\begin{align*}
\rho_{AAA} &= \rho_{CCC} = \rho_{GGG} = \rho_{TTT} & \text{4 terms} \\
\rho_{AAC} &= \rho_{AAG} = \rho_{AAT} = \cdots = \rho_{TTG} & \text{12 terms} \\
\rho_{ACA} &= \rho_{AGA} = \rho_{ATA} = \cdots = \rho_{TGT} & \text{12 terms} \\
\rho_{CAA} &= \rho_{GAA} = \rho_{TAA} = \cdots = \rho_{GTT} & \text{12 terms} \\
\rho_{ACG} &= \rho_{ACT} = \rho_{AGT} = \cdots = \rho_{CGT} & \text{24 terms}
\end{align*}$$

- For this model, the unique polynomial that detects the phylogeny of the three species has degree 3.

Definition (Cavender–Felsenstein ’87)

The generators of $I(V)$ are called phylogenetic invariants.
Computing the ideal

(Eriksson–Ranestad–Sturmfels–Sullivant ’05)

- Some generators of $I(V)$ depend only on the model chosen (not on the topology). E.g. for Jukes-Cantor they are: \( \sum p_{x_1x_2x_3} = 1 \)

\[
\begin{align*}
\rho_{AAA} &= \rho_{CCC} = \rho_{GGG} = \rho_{TTT} & &\text{4 terms} \\
\rho_{AAC} &= \rho_{AAG} = \rho_{AAT} = \cdots = \rho_{TTG} & &\text{12 terms} \\
\rho_{ACA} &= \rho_{AGA} = \rho_{ATA} = \cdots = \rho_{TGT} & &\text{12 terms} \\
\rho_{CAA} &= \rho_{GAA} = \rho_{TAA} = \cdots = \rho_{GTT} & &\text{12 terms} \\
\rho_{ACG} &= \rho_{ACT} = \rho_{AGT} = \cdots = \rho_{CGT} & &\text{24 terms}
\end{align*}
\]

- For this model, the unique polynomial that detects the phylogeny of the three species has degree 3.

Definition (Cavender–Felsenstein ’87)

The generators of $I(V)$ are called phylogenetic invariants.
Computing the ideal

(Eriksson–Ranestad–Sturmfels–Sullivant ’05)

- Some generators of $I(V)$ depend only on the model chosen (not on the topology). E.g. for Jukes-Cantor they are: $\sum p_{x_1 x_2 x_3} = 1$

  \[
  \begin{align*}
  p_{AAA} &= p_{CCC} = p_{GGG} = p_{TTT} & \text{4 terms} \\
  p_{AAC} &= p_{AAG} = p_{AAT} = \cdots = p_{TTG} & \text{12 terms} \\
  p_{ACA} &= p_{AGA} = p_{ATA} = \cdots = p_{TGT} & \text{12 terms} \\
  p_{CAA} &= p_{GAA} = p_{TAA} = \cdots = p_{GTT} & \text{12 terms} \\
  p_{ACG} &= p_{ACT} = p_{AGT} = \cdots = p_{CGT} & \text{24 terms}
  \end{align*}
  \]

- For this model, the unique polynomial that detects the phylogeny of the three species has degree 3.

Definition (Cavender–Felsenstein ’87)

The generators of $I(V)$ are called **phylogenetic invariants**.
Problem: computation of invariants

- Computational algebra software fail to compute the ideal for $\geq 4$ species! *Kimura 3-parameter, 4 species, 8002 generators like:*

[Small trees webpage]
Problem: computation of invariants

- Computational algebra software fail to compute the ideal for $\geq 4$ species! Kimura 3-parameter, 4 species, 8002 generators like:

[Small trees webpage]
Group-based models (Kimura and Jukes-Cantor): $G = \mathbb{Z}_2 \times \mathbb{Z}_2$, $A = (0, 0)$, $C = (0, 1)$, $G = (1, 0)$, $T = (1, 1)$. Substitution matrices are functions on the group:

$$S_i(g_p, g_c) = f^i(g_p - g_c), \quad g_c, g_p \in G.$$ 

$$p_{g_1, \ldots, g_m} = p(g_1, \ldots, g_m) = \frac{1}{4} \sum \prod_{e \in \text{edges}} f^e(g_p(e) - g_c(e))$$

sum over all possible values at interior nodes.

**Discrete Fourier transform** $f : G \longrightarrow \mathbb{C}$

$$\hat{f}(\chi) = \sum_{g \in G} \chi(g)f(g), \quad \chi \in \text{Hom}(G, \mathbb{C}^*) \cong G$$

Convolution: $(f_1 \ast f_2)(g) = \sum_{h \in G} f_1(h)f_2(g - h) \implies \hat{f_1} \ast \hat{f_2} = \hat{f_1} \cdot \hat{f_2}$
Group based models.

**Theorem (Evans–Speed)**

For a group-based model (i.e. Kimura 3, Kimura 2 or Jukes-Cantor) on a tree $T$, the discrete Fourier transform of the joint distribution $p(g_1, \ldots, g_n)$ has the following form

$$q(\chi_1, \ldots, \chi_m) = \prod_{e \in \text{edges}} \hat{f}_e \left( \prod_{l \in \text{leaves below } e} \chi_l \right)$$

Using a linear change of coordinates, one gets a monomial parameterization of the variety associated to the model

$$\varphi : \mathbb{C}^d \longrightarrow \mathbb{C}^{4^n}$$

$$\theta = (\hat{\theta}_1, \ldots, \hat{\theta}_d) \longmapsto (q_{AA\ldots A}, q_{AA\ldots C}, q_{AA\ldots G}, \ldots, q_{TT\ldots T})$$

so that it is a toric variety.

The ideal is generated by binomials.
**Fourier transform for Kimura 3-parameter model**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Fourier parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_e = \begin{pmatrix} a_e &amp; b_e &amp; c_e &amp; d_e \ b_e &amp; a_e &amp; d_e &amp; c_e \ c_e &amp; d_e &amp; a_e &amp; b_e \ d_e &amp; c_e &amp; b_e &amp; a_e \end{pmatrix}$</td>
<td>$P^e = \begin{pmatrix} P_A^e &amp; 0 &amp; 0 &amp; 0 \ 0 &amp; P_C^e &amp; 0 &amp; 0 \ 0 &amp; 0 &amp; P_G^e &amp; 0 \ 0 &amp; 0 &amp; 0 &amp; P_T^e \end{pmatrix}$</td>
</tr>
<tr>
<td>$P_A^e = a^e + b^e + c^e + d^e$, $P_C^e = a^e - b^e + c^e - d^e$...</td>
<td>$\Delta^3$</td>
</tr>
<tr>
<td>$P_A^e = 1$</td>
<td>$\Delta^3$</td>
</tr>
</tbody>
</table>

- **Simplex**: $\triangle^3$
  - $a_e + b_e + c_e + d_e = 1$
- **Coordinates**:
  - $p_{x_1...x_n}$
- **Simplex**: $\triangle^{4^n - 1}$
  - $\sum p_{x_1...x_n} = 1$
- **Linear invariants**:
  - $q_{g_1...g_n} = 0$ if $g_1 + \cdots + g_n \neq 0$ in $\mathbb{Z}_2 \times \mathbb{Z}_2$.
For the Kimura 3-parameter model, we are interested in
\[ V := \text{im}(\varphi) \]
where
\[ \varphi : \prod_{e \in E(T)} \mathbb{C}^4 \rightarrow \mathbb{C}^{4^{n-1}} \]
\[ (P_e^A, P_e^C, P_e^G, P_e^T)_e \mapsto (q_{x_1...x_n}) \{ x_1,...,x_n | x_1+...+x_n=0 \} \]

But \( \Delta^{4^{n-1}-1} \subset \{ q_{A...A} = 1 \} \).

**Definition**

The **Kimura variety** of the phylogenetic tree \( T \) is \( W := V \cap \{ q_{A...A} = 1 \} \).
Recursive construction of invariants

- Computational algebra software: even in Fourier parameterization, fail for $\geq 5$ leaves.

**Theorem (Sturmfels – Sullivant ’05)**

For any group-based model, they give an explicit algorithm for obtaining the generators of the ideal of phylogenetic invariants of an $n$-leaved tree from those of an unrooted tree of 3 leaves. The generators are binomials of degree $\leq 4$. 
Outline

1. Algebraic evolutionary models
2. Phylogenetic inference using algebraic geometry
3. The geometry of the Kimura variety
4. New results on simulated data
1990-1995: Biologists claim that phylogenetic invariants are not useful for phylogenetic reconstruction. Lake only used linear invariants!

In particular, a work of Huelsenbeck’95 reveals the inefficiency of Lake’s method of invariants.

(Eriksson ’05) for General Markov Model.

(C–Garcia–Sullivant ’05) for Kimura 3-parameter.
1990-1995: Biologists claim that phylogenetic invariants are not useful for phylogenetic reconstruction. Lake only used linear invariants!

In particular, a work of Huelsenbeck’95 reveals the inefficiency of Lake’s method of invariants.

(Eriksson ’05) for General Markov Model.

(C–Garcia–Sullivant ’05) for Kimura 3-parameter.
Model: Unrooted tree of 4 leaves, Kimura 3-parameters.

Naive algorithm: Take $|| \cdot ||_1$ of the evaluation of all invariants on the point given by the relative frequencies of the columns of the alignment. Obtain a score for each possible topology and choose the one with smaller score.
Model: Unrooted tree of 4 leaves, Kimura 3-parameters.

Naive algorithm: Take $\| \cdot \|_1$ of the evaluation of all invariants on the point given by the relative frequencies of the columns of the alignment. Obtain a score for each possible topology and choose the one with smaller score.

Huelsenbeck ’95: assesses the performance of different phylogenetic reconstruction methods on the following tree space.
Results on simulated data (Huelsenbeck)

Huelsenbeck’s results:

<table>
<thead>
<tr>
<th>Length 100</th>
<th>Length 500</th>
<th>Length 1000</th>
</tr>
</thead>
</table>

Lake invariants

NJ

ML

C–Fernandez-Sanchez studies on C–Garcia–Sullivant method

Length 100

Length 500

Length 1000
Why using phylogenetic invariants

1. We do not need to estimate parameters of the model.
2. The algebraic model allows species within the tree to evolve at different mutation rates
   - In the non-algebraic version, one has: \( S_i = \exp(Q \cdot t_i) \) where \( Q \) is a fixed matrix that represents the instantaneous mutation rate of all the species in the tree.
   - In the algebraic model, one is allowing different rates at each branch of the tree.
For instance using algebraic geometry one should be able to reconstruct the biologically correct tree:

(Al-Aidroos, Snir ’05) chapter 21 in ASCB book.
For instance using algebraic geometry one should be able to reconstruct the biologically correct tree:

(Al-Aidroos, Snir ’05) chapter 21 in ASCB book.
For instance using algebraic geometry one should be able to reconstruct the biologically correct tree:

(Al-Aidroos, Snir ’05) chapter 21 in ASCB book.
Simulations on non-homogeneous trees (different rate matrices)

Comparison with Neighbour joining

Comparison with Maximum Likelihood
Outline

1. Algebraic evolutionary models
2. Phylogenetic inference using algebraic geometry
3. The geometry of the Kimura variety
4. New results on simulated data
Recall that $V := \text{im}(\varphi)$ where

$$\varphi : \prod_{e \in E(T)} \mathbb{C}^4 \rightarrow \mathbb{C}^{4n-1}$$

$$\left(P_e^A, P_e^C, P_e^G, P_e^T\right) \mapsto (q_{x_1...x_n})\{x_1, ..., x_n | x_1 + ... + x_n = 0\}$$

But $\Delta^{4n-1} \subset \{q_{A...A} = 1\}$.

**Definition**

The **Kimura variety** of the phylogenetic tree $T$ is $W := V \cap \{q_{A...A} = 1\}$.

$$W = \varphi(\prod_{e \in E(T)}(\mathbb{C}^4 \cap \{P_e^A = 1\})).$$
Local complete intersection

- $V \subset \mathbb{A}^N$ algebraic variety, $\mu$ minimum number of generators of $I(V)$, then
  \[ \mu \geq \text{codim}(V) = N - \text{dim}(V) \]

- $V$ is called a complete intersection if $\mu = \text{codim}(V)$.

- Example: the Kimura variety $W \subset \mathbb{C}^{4n-1-1}$ has codimension $4^{n-1} - 6n + 8$, $n =$ number of leaves.

- For $n = 4$, the minimum number of generators of the Kimura variety is 8002 whereas its codimension is 48!

- On a neighborhood of a smooth point, any variety is a local complete intersection (i.e. it can be defined by $\text{codim}(V)$ equations).

- Goal:
  1. Find the singular points.
  2. Provide a set of local generators at non-singular points.
Local complete intersection

- \( V \subset \mathbb{A}^N \) algebraic variety, \( \mu \) minimum number of generators of \( I(V) \), then
  \[ \mu \geq \text{codim}(V) = N - \text{dim}(V) \]

- \( V \) is called a complete intersection if \( \mu = \text{codim}(V) \).

- Example: the Kimura variety \( W \subset \mathbb{C}^{4n-1-1} \) has codimension \( 4^{n-1} - 6n + 8 \), \( n = \) number of leaves.
- For \( n = 4 \), the minimum number of generators of the Kimura variety is 8002 whereas its codimension is 48!

- On a neighborhood of a smooth point, any variety is a local complete intersection (i.e. it can be defined by \( \text{codim}(V) \) equations).

- Goal:
  1. Find the singular points.
  2. Provide a set of local generators at non-singular points.
Local complete intersection

- $V \subset \mathbb{A}^N$ algebraic variety, $\mu$ minimum number of generators of $I(V)$, then

$$\mu \geq \text{codim}(V) = N - \text{dim}(V)$$

- $V$ is called a complete intersection if $\mu = \text{codim}(V)$.

- Example: the Kimura variety $W \subset \mathbb{C}^{4^{n-1}-1}$ has codimension $4^{n-1} - 6n + 8$, $n =$ number of leaves.

- For $n = 4$, the minimum number of generators of the Kimura variety is 8002 whereas its codimension is 48!

- On a neighborhood of a smooth point, any variety is a local complete intersection (i.e. it can be defined by $\text{codim}(V)$ equations).

- Goal:
  1. Find the singular points.
  2. Provide a set of local generators at non-singular points.
Case $n = 3$

The parameterization $\varphi$ is:

$$\varphi : \mathbb{C}^9$$

$$((1,P_1^1,P_1^1,P_1^1),(1,P_2^2,P_2^2,P_2^2),(1,P_3^3,P_3^3,P_3^3))$$

$$\mapsto$$

$$\mathbb{C}^{15}$$

$$\{q_{xyz}=P_1^1P_2^2P_3^3\} \{x+y+z=0\}$$

Let $H = (\mathbb{Z}_2 \times \mathbb{Z}_2, *)$ and let $(\varepsilon, \delta) \in H$ act on $\mathbb{C}^9$ sending $(P_1^1, P_2^2, P_3^3)$ to

$$((1,\varepsilon P_1^1,\delta P_1^1,\varepsilon \delta P_1^1),(1,\varepsilon P_2^2,\delta P_2^2,\varepsilon \delta P_2^2),(1,\varepsilon P_3^3,\delta P_3^3,\varepsilon \delta P_3^3)).$$

E.g. $(\varepsilon, \delta) = (1, -1)$, in probability parameters the group action permutes $A \leftrightarrow C$ and $G \leftrightarrow T$.

**Proposition**

The Kimura variety $W = \overline{\text{im}(\varphi)}$ on $T_3$ is the affine GIT quotient $\mathbb{C}^9 // H$. 

M. Casanellas (UPC)  
Using algebraic geometry ...  
IMA, March 2007  29 / 36
Arbitrary \(n\)

\(T\) unrooted \(n\)-leaved tree
(\(2n - 3\) edges, \(n - 2\) interior nodes)

Extending the action of \(H\) to the \(n - 2\) interior nodes of \(T\), we have

**Theorem**

*The Kimura variety \(W\) is isomorphic to the affine GIT quotient*

\[
(C^3)^{2n-3} / / H^{n-2}
\]

**Corollary**

- \(W = \text{im}(\varphi)\), no closure needed.
- \(|\varphi^{-1}(q)| \leq 4^{n-2}\) and there is just one preimage with biological meaning, \(\forall q \in W\).
- We are able to determine the singular points. In particular, biologically meaningful points \(q \in W_+ := \varphi(\prod \Delta^3_+)\) are not singular.
**Theorem**

The Kimura variety $W$ is isomorphic to the affine GIT quotient

$$(\mathbb{C}^3)^{2n-3} // H^{n-2}$$

**Corollary**

- $W = im(\varphi)$, no closure needed.
- $|\varphi^{-1}(q)| \leq 4^{n-2}$ and there is just one preimage with biological meaning, $\forall q \in W$.
- We are able to determine the singular points. In particular, biologically meaningful points $q \in W_+ := \varphi(\prod \Delta_+^3)$ are not singular.
Arbitrary $n$

$T$ unrooted $n$-leaved tree
(2$n$ − 3 edges, $n$ − 2 interior nodes)

Extending the action of $H$ to the $n$ − 2 interior nodes of $T$, we have

**Theorem**

The Kimura variety $W$ is isomorphic to the affine GIT quotient

$$(\mathbb{C}^3)^{2n-3} /\!/ H^{n-2}$$

**Corollary**

- $W = \text{im}(\varphi)$, no closure needed.
- $|\varphi^{-1}(q)| \leq 4^{n-2}$ and there is just one preimage with biological meaning, $\forall q \in W$.

- We are able to determine the singular points. In particular, biologically meaningful points $q \in W_+ := \varphi(\prod \Delta_+^3)$ are not singular.
Arbitrary $n$

$T$ unrooted $n$-leaved tree
$(2n - 3 \text{ edges, } n - 2 \text{ interior nodes})$

Extending the action of $H$ to the $n - 2$ interior nodes of $T$, we have

**Theorem**

The Kimura variety $W$ is isomorphic to the affine GIT quotient

$$(\mathbb{C}^3)^{2n-3} \sslash H^{n-2}$$

**Corollary**

$W = \text{im}(\varphi)$, no closure needed.

$|\varphi^{-1}(q)| \leq 4^{n-2}$ and there is just one preimage with biological meaning, $\forall q \in W$.

We are able to determine the singular points. In particular, biologically meaningful points $q \in W_+ := \varphi(\prod \Delta^3_+)$ are not singular.
In Fourier parameters:

\[
\Delta^3_+ =
\]

In probability parameters this is transformed into:

\[
S = \begin{pmatrix}
A & c & d & d \\
C & b & a & d \\
G & c & d & a \\
T & d & c & b \\
\end{pmatrix}
\]

\[
\begin{align*}
A + B + C + D &= 1 \\
A + B - C - D &> 0 \\
A - B + C - D &> 0 \\
A - B - C + D &> 0
\end{align*}
\]
In Fourier parameters:

\[ \Delta^3_+ = \]

In probability parameters this is transformed into:

\[
S = \begin{pmatrix}
A & a & b & c & d \\
C & b & a & d & c \\
G & c & d & a & b \\
T & d & c & b & a
\end{pmatrix}
\]

\[
a + b + c + d = 1
\]

\[
a + b > 1/2
\]

\[
a + c > 1/2
\]

\[
a + d > 1/2
\]
Local complete intersection

Case $n = 3$

- (Sturmfels–Sullivant ’05): $I(V)$ is minimally generated by 16 cubics and 18 quartics.

Lemma

The following six quartics

\[
\begin{align*}
q_{\text{AAA}}q_{\text{ATT}}q_{\text{TCG}}q_{\text{TGC}} - q_{\text{ACC}}q_{\text{AGG}}q_{\text{TAT}}q_{\text{TTA}}, & \quad q_{\text{CCA}}q_{\text{CTG}}q_{\text{TAT}}q_{\text{TGC}} - q_{\text{CAC}}q_{\text{CGT}}q_{\text{TCG}}q_{\text{TTA}}, \\
q_{\text{AGG}}q_{\text{ATT}}q_{\text{CAC}}q_{\text{CCA}} - q_{\text{AAA}}q_{\text{ACC}}q_{\text{CGT}}q_{\text{CTG}}, & \quad q_{\text{ACC}}q_{\text{ATT}}q_{\text{GAG}}q_{\text{GGA}} - q_{\text{AAA}}q_{\text{AGG}}q_{\text{GCT}}q_{\text{GTC}}, \\
q_{\text{CAC}}q_{\text{CTG}}q_{\text{GCT}}q_{\text{GGA}} - q_{\text{CCA}}q_{\text{CGT}}q_{\text{GAG}}q_{\text{GTC}}, & \quad q_{\text{GGA}}q_{\text{GTC}}q_{\text{TAT}}q_{\text{TGC}} - q_{\text{GAG}}q_{\text{GCT}}q_{\text{TCG}}q_{\text{TTA}}
\end{align*}
\]

generate a local complete intersection that defines $W$ at each point $q \in W_+$.

It does not depend on the point $q$. 
Proof.

1. $J = (f_1 \ldots, f_6)$ defines a variety $X$ containing $W$ and both have the same dimension.

2. The points $q \in W_+$ are smooth points of $X$ and $W$.

3. Both varieties coincide locally.
Local complete intersection

Arbitrary $n$

- quartics:
- quadrics: $Q$, non-redundant $2 \times 2$ minors from flattening

Theorem

$J_3 \cup J_{n-1} \cup Q$ generate a local complete intersection that defines $W$ at the biologically meaningful points $q \in W_+$. 
Local complete intersection

Theorem

\[ J_3 \cup J_{n-1} \cup Q \text{ generate a local complete intersection that defines } W \text{ at the biologically meaningful points } q \in W_+. \]

- quartics:
  - quadrics: \( Q \), non-redundant \( 2 \times 2 \) minors from flattening

Arbitrary \( n \)
Local complete intersection

- quartics: $J_3$
- quadrics: $Q$, non-redundant $2 \times 2$ minors from flattening

**Theorem**

$J_3 \cup J_{n-1} \cup Q$ generate a local complete intersection that defines $W$ at the biologically meaningful points $q \in W_+$. 

M. Casanellas (UPC) Using algebraic geometry ... IMA, March 2007 34 / 36
Local complete intersection

Arbitrary $n$

- quartics: $J_3$
- quadrics: $Q$, non-redundant $2 \times 2$ minors from flattening

**Theorem**

$J_3 \cup J_{n-1} \cup Q$ generate a local complete intersection that defines $W$ at the biologically meaningful points $q \in W_+$. 

M. Casanellas (UPC)  
Using algebraic geometry ...  
IMA, March 2007  

34 / 36
Local complete intersection

Arbitrary $n$

- quartics: $J_3 \cup J_{n-1}$
- quadrics: $Q$, non-redundant $2 \times 2$ minors from flattening

Theorem

$J_3 \cup J_{n-1} \cup Q$ generate a local complete intersection that defines $W$ at the biologically meaningful points $q \in W_+$. 
Local complete intersection

**Theorem**

\[ J_3 \cup J_{n-1} \cup Q \text{ generate a local complete intersection that defines } W \text{ at the biologically meaningful points } q \in W_+. \]
Local complete intersection

Arbitrary $n$

- quartics: $J_3 \cup J_{n-1}$
- quadrics: $Q$, non-redundant $2 \times 2$ minors from flattening

**Theorem**

$J_3 \cup J_{n-1} \cup Q$ generate a local complete intersection that defines $W$ at the biologically meaningful points $q \in W_+$. 
Outline

1. Algebraic evolutionary models
2. Phylogenetic inference using algebraic geometry
3. The geometry of the Kimura variety
4. New results on simulated data
New results on simulated data

• Instead of all generators, 48 polynomials that generate locally

Length 100  Length 500  Length 1000
New results on simulated data

4 leaves

- Instead of all generators, 48 polynomials that generate locally

![Length 100](image1)
![Length 500](image2)
![Length 1000](image3)

![RP Length 100](image4)
![Length 500](image5)
![RP Length 1000](image6)
New results on simulated data

4 leaves

- Instead of all generators, 48 polynomials that generate locally
  - Length 100
  - Length 500
  - Length 1000

- (Eriksson–Yao’07) Machine learning approach, 52 invariants that perform better on this parameter space.