Finite phylogenetic complexity and combinatorics of tables
joint work with M. Michałek

Emanuele Ventura
Max Planck Institute for Mathematics in the Sciences, Leipzig

June 14th, MEGA, Nice
**Phylogenetics** is the part of computational biology that models evolution and describes mutations in this process.
Trees of species
The ultimate goal is to have an *atlas* of the origin of Life.

A tree of life, showing the relationship between species whose genomes had been sequenced as of 2006. The very centre represents the last universal ancestor of all life on earth. The different colors represent the three domains of life: pink represents eukaryota (animals, plants and fungi); blue represents bacteria; and green represents archaea. Note the presence of homo sapiens (humans) second from the rightmost edge of the pink segment. Click [here](/content/sites/plus.maths.org/files/articles/2016/evolution/real_tree_large.svg) to see a larger version of this picture.
Applications of Phylogenetics

Phylogenetics is a fertile subject and applications of it have appeared in:

- algebraic geometry
- combinatorics
- representation theory
- humanities
Introduction
Equations and phylogenetic complexity
Kimura 3-parameter model

Tree models: an example

Evolution is *modeled* with rooted trees and transition matrices along its edges. The vertices are the species.

**Example**

The root $r$ is the *ancestor*. $\Pr(r = 0) = \sigma_0$ and $\Pr(r = 1) = \sigma_1$. The matrices on the edges $a$ and $b$ are

$$A = \begin{bmatrix} a_0 & a_1 \\ a_1 & a_0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} b_0 & b_1 \\ b_1 & b_0 \end{bmatrix},$$

$a_0 = \Pr(l_0 = 0|r = 0) = \Pr(l_0 = 1|r = 1)$, $a_1 = \Pr(l_0 = 0|r = 1) = \Pr(l_0 = 1|r = 0)$. Same for $b_0, b_1$. 

Tree models: an example, continued

Example

Parameters of the tree model are $\sigma_0, \sigma_1, a_0, a_1, b_0, b_1$. The probability distribution at the leaves $l_0, l_1$ is this matrix with the possible states 00, 01, 10, 11:

\[
\begin{bmatrix}
\sigma_0 a_0 b_0 + \sigma_1 a_1 b_1 & \sigma_0 a_0 b_1 + \sigma_1 a_1 b_0 \\
\sigma_0 a_1 b_0 + \sigma_1 a_0 b_1 & \sigma_0 a_1 b_1 + \sigma_1 a_0 b_0
\end{bmatrix}.
\]

This matrix can be understood as a projective morphism $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \to \mathbb{P}^3(00,01,10,11)$.

A phylogenetic tree model is encoded by a projective variety. Generators of ideals of phylogenetic varieties are the phylogenetic invariants, introduced by Cavender and Felsenstein, and Lake.
Definition

A phylogenetic tree model consists of:

- $\mathcal{T} = \text{directed rooted tree}$
- $X_v = \text{random variable corresponding to vertex } v \text{ in } \mathcal{T}$
- $S = \text{finite set of states of } X_v \text{ for all } v \text{ in } \mathcal{T}$
- $W = \bigoplus_{s \in S} \mathbb{C}_s$
- To each vertex $v$ in $\mathcal{T}$ we attach a vector space $W_v$ isomorphic to $W$

Definition

A model is a choice of a subspace $\tilde{W} \subseteq \text{End}(W)$ to be associated to each edge in $\mathcal{T}$. 
(i) The *Jukes-Cantor model*: $S$ has four elements and the matrices are

$$\begin{bmatrix}
a & b & b & b \\
b & a & b & b \\
b & b & a & b \\
b & b & b & a \\
\end{bmatrix}$$

(ii) The *Cavender-Farris-Neyman model*: $S$ has two elements and the matrices are

$$\begin{bmatrix}
a & b \\
b & a \\
\end{bmatrix}$$

(iii) The *Kimura 2-parameter model*: $S$ has four elements and the matrices are

$$\begin{bmatrix}
a & b & c & b \\
b & a & b & c \\
c & b & a & b \\
b & c & b & a \\
\end{bmatrix}$$
Tree models, III

(iv) The *Kimura 3-parameter model*: $S$ has four elements and the matrices are

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

(v) The *general Markov model*: all matrices in $\text{End}(W)$ and the number of states is arbitrary.
Introduction
Equations and phylogenetic complexity
Kimura 3-parameter model

Group-based models

**Definition (Evans-Speed, Sturmfels-Sullivant)**

Let $G$ be a group. A group-based model associated to $G$ is a model where the set of states $S = G$, and the probability of transitioning from an element $g_i \in G$ to another $g_j \in G$ only depends on $g_i - g_j$.

**Example**

The Kimura 3-parameter model is the group-based model for $G = \mathbb{Z}_2 \times \mathbb{Z}_2$. The set of states is $S = \{A, G, C, T\}$, the four nucleobases. This reflects the pairing between purines ($A, G$) and pyrimidines ($C, T$).
Varieties for group-based models and flows

The variety encoding a group-based model over the group $G$ on the star with $n$ leaves $K_{1,n}$ is a toric variety, denoted $X(G, K_{1,n})$.

**Definition**

Let $G$ be a finite abelian group and $n \in \mathbb{N}$. A flow is a sequence of $n$ elements of $G$ summing up to $0 \in G$, the neutral element of $G$. The collection of flows is a group via the coordinate-wise multiplication.

**Theorem (Sturmfels-Sullivant, Michałek)**

Flows are in bijection with the monomials parameterizing $X(G, K_{1,n})$. 
**Example:** $X(G, K_{1,3})$

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Let $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ and $n = 3$.</td>
</tr>
<tr>
<td>Identify $A = (0, 0)$, $G = (0, 1)$, $C = (1, 0)$, $T = (1, 1)$.</td>
</tr>
<tr>
<td>The ideal of $X(G, K_{1,3})$ sits in $\mathbb{C}[q_{AAA}, q_{AGG}, \ldots, q_{TTA}]$ and it is the kernel of the monomial map:</td>
</tr>
<tr>
<td>$q(g_1, g_2, g_3) \mapsto x_{g_1} y_{g_2} z_{g_3}$.</td>
</tr>
<tr>
<td>This toric ideal is generated by 16 binomials of degree 3 and 18 binomials of degree 4. This is a 9-dimensional toric variety in $\mathbb{P}^{15}$.</td>
</tr>
</tbody>
</table>
As usual in the toric setting, the binomial relations among the flows are the equations of the variety $X(G, K_{1,n})$. These binomial relations are linear relations of the vertices of the polytope of $X(G, K_{1,n})$.

Example

Let $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ and $n = 3$. Identify $A = (0,0)$, $G = (0,1)$, $C = (1,0)$, $T = (1,1)$. The binomial of degree 3

$$q_{AAA}q_{CTG}q_{TGC} - q_{AGG}q_{CAC}q_{TTA}$$

is in the ideal of $X(G, K_{1,3})$. Encode it as follows:

$$T_0 - T_1 = \begin{bmatrix} A & A & A \\ C & T & G \\ T & G & C \end{bmatrix} - \begin{bmatrix} A & G & G \\ C & A & C \\ T & T & A \end{bmatrix}.$$
Introduction
Equations and phylogenetic complexity
Kimura 3-parameter model

Binomials and compatible tables

**Lemma**

A binomial belongs to $I(X(G, K_{1,n}))$ if and only if the two tables are compatible: For each $i$, the $i$-th column of $T_0$ and the $i$-th column of $T_1$ are equal as multisets.

**Remark**

*Generation*: to generate a binomial in the ideal (given by the pair $T_0$, $T_1$) by binomials in the ideal of degree at most $k$, we select a subset of rows in $T_0$ of cardinality at most $k$ and replace it with a compatible set of rows, repeating this procedure until $T_0$, $T_1$ are equal.
Example of compatible tables, I

Example

For $G = (\mathbb{Z}_2, +)$ and $n = 6$ consider the following two compatible tables:

$$T_0 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad T_1 = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}.$$  

Note that the red subtable of $T_0$ is compatible with the table

$$T' = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$
Example of compatible tables, II

Example

Exchange them obtaining:

\[ \tilde{T}_0 = \begin{bmatrix}
0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 & 0
\end{bmatrix}. \]

\( T_0 \) and \( \tilde{T}_0 \) are compatible. Brown subtable of \( \tilde{T}_0 \) is compatible with the table

\[ T'' = \begin{bmatrix}
1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 1
\end{bmatrix}. \]

We exchange them obtaining \( T_1 \). We have a sequence of tables \( T_0 \rightsquigarrow \tilde{T}_0 \rightsquigarrow T_1 \): we started from a degree 3 binomial given by the pair \( T_0, T_1 \) and we generated it using degree 2 binomials. These are quadratic moves.
**Phylogenetic complexity**

**Definition (Sturmfels-Sullivant)**

Let $K_{1,n}$ be the star with $n$ leaves, and let $\phi(G, K_{1,n})$ be the maximal degree of a generator in a minimal generating set of $I(X(G, K_{1,n}))$. Let $\phi(G, n) = \phi(G, K_{1,n})$. The phylogenetic complexity $\phi(G)$ of $G$ is $\sup_{n \in \mathbb{N}} \{\phi(G, n)\}$.

**Theorem (Sturmfels-Sullivant)**

The equations of $X(G, T)$, for any tree $T$, can be recovered by the ones of stars $K_{1,n}$, possibly with additional quadrics.

**Remark**

Phylogenetic complexity gives an upper bound of the minimal degrees of phylogenetic invariants of the ideals of $X(G, K_{1,n})$, and hence for all ideals of $X(G, T)$.
Finiteness of phylogenetic complexity

**Theorem**

For any finite abelian group $G$, the phylogenetic complexity $\phi(G)$ is finite.

Sturmfels and Sullivant conjectured a much stronger statement:

**Conjecture**

For any finite abelian group $G$, the phylogenetic complexity satisfies $\phi(G) \leq |G|$.

A lot of related results concern finiteness of degrees and of equations of infinite families of varieties: work of Aschenbrenner, Draisma, Eggermont, Hillar, Krone, Kuttler, Leykin, Martin del Campo, Sam, Snowden, Sullivant et al.
Open questions

Questions

- *In this approach we do not provide an explicit bound (maybe not even polynomial in $|G|$?) and the conjectural one is $|G|$. How to get such a bound?*

- *What about non-abelian groups?*

- *It seems crucial to first understand the case of a tripod, i.e., $K_{1,3}$. Is it true that for any abelian group $G$, $\phi(G, 3) \leq |G|$?*
We made our approach on tables more explicit for the case of the Kimura 3-parameter model. Recall that:

**Remark**

*This is a group-based model for $G = \mathbb{Z}_2 \times \mathbb{Z}_2$."

**Theorem (Conjectured by Sturmfels-Sullivant)**

*The phylogenetic complexity of the Kimura 3-parameter model $\phi(\mathbb{Z}_2 \times \mathbb{Z}_2)$ is 4.*
Thank you