

Geometry of phylogenetics

W. Buczyńska, J.A. Wiśniewski

Institute of Mathematics, Warsaw University, Poland

phylogenetics

Phylogenetics: reconstructing historical relation between species by analyzing their present features and putting their common ancestors in a diagram which forms a tree. [e.g. Häckel, 1866]













[lemma] There is a bijection between the set of sockets and networks, that is for every socket σ there exists a unique network $\mu(\sigma)$ whose end points are in σ

[lemma] There is a bijection between the set of sockets and networks, that is for every socket σ there exists a unique network $\mu(\sigma)$ whose end points are in σ

For every edge $e \in \mathcal{E}$ we consider a \mathbb{P}_e^1 with homogeneous coordinates $[y_0^e, y_1^e]$. Moreover consider a projective space \mathbb{P}_{Σ} of dimension $2^d - 1$ with homogeneous coordinates $[z_{\sigma}]$ indexed by sockets of \mathcal{T} . [lemma] There is a bijection between the set of sockets and networks, that is for every socket σ there exists a unique network $\mu(\sigma)$ whose end points are in σ

Define rational map $\prod_{e \in \mathcal{E}} \mathbb{P}^1_e \to \mathbb{P}_{\Sigma}$ such that

$$z_{\sigma} = \prod_{e \in \mu(\sigma)} y_1^e \cdot \prod_{e \notin \mu(\sigma)} y_0^e$$

The model of the tree, $X(\mathcal{T}) \subset \mathbb{P}_{\Sigma}$, is the closure of the image of this map, $\dim X(\mathcal{T}) = 2d - 1$.

Leaves of \mathcal{T} are labeled by numbers $1, \ldots, d+1$ and sockets are denoted by 0/1 of length d+1. Leaves of \mathcal{T} are labeled by numbers $1, \ldots, d+1$ and sockets are denoted by 0/1 of length d+1. Edges are labeled by letters. Leaves of \mathcal{T} are labeled by numbers $1, \ldots, d+1$ and sockets are denoted by 0/1 of length d+1. Tripod tree model:

$$\mathbb{P}_{a}^{1} \times \mathbb{P}_{b}^{1} \times \mathbb{P}_{c}^{1} \to \mathbb{P}^{3}$$

$$z_{000} = y_{0}^{a} y_{0}^{b} y_{0}^{c} \quad z_{110} = y_{1}^{a} y_{1}^{b} y_{0}^{c}$$

$$z_{101} = y_{1}^{a} y_{0}^{b} y_{1}^{c} \quad z_{011} = y_{0}^{a} y_{1}^{b} y_{1}^{c}$$



Leaves of \mathcal{T} are labeled by numbers $1, \ldots, d+1$ and sockets are denoted by 0/1 of length d+1. Four leaf tree model in \mathbb{P}^7

 $\begin{aligned} z_{0000} &= y_0^a y_0^b y_0^c y_0^d y_0^e \quad z_{1111} = y_1^a y_1^b y_0^c y_1^d y_1^e \quad 2 \qquad 3 \\ z_{1100} &= y_1^a y_1^b y_0^c y_0^d y_0^e \quad z_{0011} = y_0^a y_0^b y_0^c y_1^d y_1^e \quad b < c < d \\ z_{1010} &= y_1^a y_0^b y_1^c y_1^d y_0^e \quad z_{1001} = y_1^a y_0^b y_1^c y_0^d y_1^e \quad a < c < d \\ z_{0110} &= y_0^a y_1^b y_1^c y_1^d y_0^e \quad z_{0101} = y_0^a y_1^b y_1^c y_0^d y_1^e \quad 1 \qquad 4 \end{aligned}$

Leaves of \mathcal{T} are labeled by numbers $1, \ldots, d+1$ and sockets are denoted by 0/1 of length d+1. Therefore $X(\succ) \simeq \mathbb{P}^3$ and $X(\succ)$ is a complete intersection in \mathbb{P}^7 :

Fix a root r in tree \mathcal{T} : this implies a partial order < on the set of vertexes $\mathcal{V} = \mathcal{L} \cup \mathcal{N}$. To each vertex $v \in \mathcal{V}$ assign a random variable ξ_v which takes value in $\{\alpha_1, \alpha_2\}$.

Fix a root r in tree \mathcal{T} : this implies a partial order < on the set of vertexes $\mathcal{V} = \mathcal{L} \cup \mathcal{N}$. To each vertex $v \in \mathcal{V}$ assign a random variable ξ_v which takes value in $\{\alpha_1, \alpha_2\}$. Variables ξ_v determine a Markov process on \mathcal{T} if (intuitively) the value of ξ_v depends only on the value of ξ_u , where u is the node immediately preceding v.

Fix a root r in tree \mathcal{T} : this implies a partial order < on the set of vertexes $\mathcal{V} = \mathcal{L} \cup \mathcal{N}$. To each vertex $v \in \mathcal{V}$ assign a random variable ξ_v which takes value in $\{\alpha_1, \alpha_2\}$. For each edge $e = \langle u, v \rangle$ bounded by vertexes u < v define the transition matrix A^e :

$$A_{ij}^e = P(\xi_v = \alpha_j | \xi_u = \alpha_i)$$

Fix a root r in tree \mathcal{T} : this implies a partial order < on the set of vertexes $\mathcal{V} = \mathcal{L} \cup \mathcal{N}$. To each vertex $v \in \mathcal{V}$ assign a random variable ξ_v which takes value in $\{\alpha_1, \alpha_2\}$. For each edge $e = \langle u, v \rangle$ bounded by vertexes u < v define the transition matrix A^e :

$$A_{ij}^e = P(\xi_v = \alpha_j | \xi_u = \alpha_i)$$

and set the probability of the variable ξ_r at the root: $P_i^r = P(\xi_r = \alpha_i)$

For a Markov process on a rooted tree $\ensuremath{\mathcal{T}}$ as above

For a Markov process on a rooted tree \mathcal{T} as above and any function $\mathcal{V} \ni v \to \rho(v) \in \{1, 2\}$

$$P(\bigwedge_{v\in\mathcal{V}}\xi_v=\alpha_{\rho(v)})=P^r_{\rho(r)}\cdot\prod_{e=\langle u,v\rangle\in\mathcal{E}}A^e_{\rho(u)\rho(v)}$$

For a Markov process on a rooted tree \mathcal{T} as above and any function $\mathcal{L} \ni v \to \rho(v) \in \{1, 2\}$

$$P(\bigwedge_{v \in \mathcal{L}} \xi_v = \alpha_{\rho(v)}) = \sum_{\widehat{\rho}} P^r_{\widehat{\rho}(r)} \cdot \prod_{e = \langle u, v \rangle \in \mathcal{E}} A^e_{\widehat{\rho}(u)\widehat{\rho}(v)}$$

where the sum is taken over all $\hat{\rho}$: $\mathcal{V} \rightarrow \{1, 2\}$ which extend ρ .

For a Markov process on a rooted tree \mathcal{T} as above and any function $\mathcal{L} \ni v \to \rho(v) \in \{1, 2\}$

$$P(\bigwedge_{v \in \mathcal{L}} \xi_v = \alpha_{\rho(v)}) = \sum_{\widehat{\rho}} P^r_{\widehat{\rho}(r)} \cdot \prod_{e = \langle u, v \rangle \in \mathcal{E}} A^e_{\widehat{\rho}(u)\widehat{\rho}(v)}$$

where the sum is taken over all $\hat{\rho} : \mathcal{V} \to \{1, 2\}$ which extend ρ . Phylogenetics: understand the shape of \mathcal{T} by looking at the distribution of $P(\bigwedge_{v \in \mathcal{L}} \xi_v = \alpha_{\rho(v)}).$

Phylogenetics wants to understand the locus of possible probability values of a Markov process on a fixed tree \mathcal{T} :

$$\begin{aligned} \mathcal{X}(\mathcal{T}) &:= \\ \{\zeta_{\rho} = P(\bigwedge_{v \in \mathcal{L}} \xi_{v} = \alpha_{\rho(v)}) : A^{e}_{ij}, P^{r}_{i} \text{ are arbitrary} \} \\ \text{in the simplex with coordinates } \zeta_{\rho} \text{ where } \zeta_{\rho} \geq 0, \\ \sum_{\rho} \zeta_{\rho} = 1. \end{aligned}$$

tree → variety, Markov view

We deal with a very special case, we assume:

tree → variety, Markov view

We deal with a very special case, we assume: • the root distribution is uniform, $P_1^r = P_2^r$

We deal with a very special case, we assume: • the root distribution is uniform, $P_1^r = P_2^r$

• the transition matrices are symmetric:

$$A_{12}^e = A_{21}^e, \ A_{11}^e = A_{22}^e$$

- the root distribution is uniform, $P_1^r = P_2^r$
- the transition matrices are symmetric:

$$A_{12}^e = A_{21}^e, \ A_{11}^e = A_{22}^e$$

note: these assumptions are very special but then incidently we have

We deal with a very special case, we assume: • the root distribution is uniform, $P_1^r = P_2^r$ • the transition matrices are symmetric:

$$A_{12}^e = A_{21}^e, \ A_{11}^e = A_{22}^e$$

[theorem: Sturmfels, Sullivant] Then after suitable change of coordinates (and identifying spaces) the varieties $\mathcal{X}(\mathcal{T})$ and $X(\mathcal{T})$ coincide.

On \mathbb{P}^3 with homogeneous coordinates $[z_{000}, z_{110}, z_{101}, z_{011}]$ take three actions of \mathbb{C}^* whose weights are determined by socket 0/1 sequence:

 $\lambda_1(t)[z_{000}, z_{110}, z_{101}, z_{011}] = [z_{000}, tz_{110}, tz_{101}, z_{011}]$

On \mathbb{P}^3 with homogeneous coordinates $[z_{000}, z_{110}, z_{101}, z_{011}]$ take three actions of \mathbb{C}^* whose weights are determined by socket 0/1 sequence:

 $\lambda_1(t)[z_{000}, z_{110}, z_{101}, z_{011}] = [z_{000}, tz_{110}, tz_{101}, z_{011}]$

Trivalent trees are built from tripods by identifying edges of leaves:

On \mathbb{P}^3 with homogeneous coordinates $[z_{000}, z_{110}, z_{101}, z_{011}]$ take three actions of \mathbb{C}^* whose weights are determined by socket 0/1 sequence:

 $\lambda_1(t)[z_{000}, z_{110}, z_{101}, z_{011}] = [z_{000}, tz_{110}, tz_{101}, z_{011}]$

Take quotient $\mathbb{P}_a^3 \times \mathbb{P}_b^3 / (\lambda_{3a} \cdot \lambda_{3b}^{-1})$

 $([z_{000}^{a}, z_{110}^{a}, z_{101}^{a}, z_{011}^{a}], [z_{000}^{b}, z_{110}^{b}, z_{101}^{b}, z_{011}^{b}]) \rightarrow \\ [z_{000}^{a} z_{000}^{b}, z_{000}^{a} z_{110}^{b}, z_{110}^{a} z_{000}^{b}, z_{110}^{a} z_{101}^{b}, z_{101}^{a} z_{101}^{b}, z_{101}^{a} z_{101}^{b}, z_{101}^{a} z_{101}^{b}, z_{101}^{a} z_{101}^{b}, z_{101}^{a} z_{101}^{b}, z_{101}^{a} z_{011}^{b}]) \rightarrow \\ z_{101}^{a} z_{011}^{b}, z_{011}^{a} z_{101}^{b}, z_{011}^{a} z_{011}^{b}]$

On \mathbb{P}^3 with homogeneous coordinates $[z_{000}, z_{110}, z_{101}, z_{011}]$ take three actions of \mathbb{C}^* whose weights are determined by socket 0/1 sequence:

 $\lambda_1(t)[z_{000}, z_{110}, z_{101}, z_{011}] = [z_{000}, tz_{110}, tz_{101}, z_{011}]$

The variety $X(\mathcal{T})$ is obtained as a quotient of product of \mathbb{P}^3 indexed by inner nodes by a torus identifying legs of tripods to inner edges of the tree.

tree \rightarrow **variety**, toric view

In toric geometry \mathbb{P}^3 can be viewed as a tetrahedron inscribed in a cube with the three principal directions of projection representing respective \mathbb{C}^* actions.



tree \rightarrow **variety, toric view**

In toric geometry \mathbb{P}^3 can be viewed as a tetrahedron inscribed in a cube with the three principal directions of projection representing respective \mathbb{C}^* actions.



Thus toric varieties associated to trees can be viewed as the fiber products of such tetrahedra.
tree \rightarrow **variety**, toric view

In toric geometry \mathbb{P}^3 can be viewed as a tetrahedron inscribed in a cube with the three principal directions of projection representing respective \mathbb{C}^* actions.



Thus toric varieties associated to trees can be viewed as the fiber products of such tetrahedra. So Hilbert-Ehrhart polynomial of $X(\mathcal{T})$ can be computed effectively.

tree \rightarrow **variety, toric view**

In toric geometry \mathbb{P}^3 can be viewed as a tetrahedron inscribed in a cube with the three principal directions of projection representing respective \mathbb{C}^* actions.



Thus toric varieties associated to trees can be viewed as the fiber products of such tetrahedra.

Surprise: Hilbert-Ehrhart polynomial does not depend on the shape of $X(\mathcal{T})$.

For a positive integer n let $[n] = \{0, ..., n\}$. Function $f : [n] \to \mathbb{Z}$ is symmetric if for every $k \in [n]$ it holds f(k) = f(n - k). By $\mathbf{1} : [n] \to \mathbb{Z}$ denote the unit function. If $f_1 f_2 : [n] \to \mathbb{Z}$ are symmetric functions then we define their symmetric product $f_1 \star f_2 : [n] \to \mathbb{Z}$ such that for $k \le n/2$:

$$(f_1 \star f_2)(k) = 2 \cdot \left(\sum_{i=0}^{k-1} \sum_{j=0}^{i} f_1(i) f_2(k+i-2j) \right) \\ + \left(\sum_{i=k}^{n-k} \sum_{j=0}^{k} f_1(i) f_2(k+i-2j) \right)$$

geometric interpretation of *****



Consider the simplex Δ as in the picture $(f_1 \star f_2)(k)$ is equal to the sum of products of f_1 and f_2 counted over points of lattice spanned by Δ in k-th slice of $n\cdot \Delta$ (1+1)(k) = (k+1)(n-k+1)is the number of lattice points in k-th slice of $n \cdot \Delta$ and thus ★ can be used to compute

Hilbert-Ehrhart polynomial

travel trough $6 \cdot \Delta$



travel trough $6 \cdot \Delta$



travel trough $6 \cdot \Delta$



travel trough $6 \cdot \Delta$



travel trough $6 \cdot \Delta$



travel trough $6 \cdot \Delta$



travel trough $6 \cdot \Delta$



• \star is commutative, $f_1 \star f_2 = f_2 \star f_1$

- \star is commutative, $f_1 \star f_2 = f_2 \star f_1$
- \star is usually non-associative, i.e. $(f_1 \star f_2) \star f_3 \neq f_1 \star (f_2 \star f_3)$

- \star is commutative, $f_1 \star f_2 = f_2 \star f_1$
- \star is usually non-associative, i.e. $(f_1 \star f_2) \star f_3 \neq f_1 \star (f_2 \star f_3)$
- however, [theorem] If Ω is the smallest set of functions closed under * and containing 1 then * is associative within Ω

- \star is commutative, $f_1 \star f_2 = f_2 \star f_1$
- \star is usually non-associative, i.e. $(f_1 \star f_2) \star f_3 \neq f_1 \star (f_2 \star f_3)$
- however, [theorem] If Ω is the smallest set of functions closed under * and containing 1 then * is associative within Ω



Recall that leaves of \mathcal{T} can be labeled by numbers $1, \ldots, d+1$ or, equivalently, given d+1points we can make them leaves of a (non-unique) tree \mathcal{T} .

Recall that leaves of \mathcal{T} can be labeled by numbers $1, \ldots, d+1$ or, equivalently, given d+1points we can make them leaves of a (non-unique) tree \mathcal{T} .



Recall that leaves of \mathcal{T} can be labeled by numbers $1, \ldots, d+1$ or, equivalently, given d+1points we can make them leaves of a (non-unique) tree \mathcal{T} . Thus, all varieties representing different labeled

trees can be embedded in a fixed \mathbb{P}_{Σ}

Recall that leaves of \mathcal{T} can be labeled by numbers $1, \ldots, d+1$ or, equivalently, given d+1points we can make them leaves of a (non-unique) tree \mathcal{T} .

Thus, all varieties representing different labeled trees can be embedded in a fixed \mathbb{P}_{Σ} These varieties can be non-isomorphic (one can

check it), however they are in the same connected component of the Hilbert scheme of \mathbb{P}_{Σ} , that is

Recall that leaves of \mathcal{T} can be labeled by numbers $1, \ldots, d+1$ or, equivalently, given d+1points we can make them leaves of a (non-unique) tree \mathcal{T} . Thus, all varieties representing different labeled trees can be embedded in a fixed \mathbb{P}_{Σ} [theorem] $X(\mathcal{T}_1)$ can be deformed to $X(\mathcal{T}_2)$ if only \mathcal{T}_1 and \mathcal{T}_2 have the same number of leaves.

Translate the original problem into toric geometry

Translate the original problem into toric geometry

tree

Translate the original problem into toric geometry

tree

variety

Translate the original problem into toric geometry

tree

polytope variety

Translate the original problem into toric geometry

tree polytope variety

understand the basic objects

Translate the original problem into toric geometry





polytope variety

polytope

Translate the original problem into toric geometry

variety

 \mathbb{P}^3





Translate the original problem into toric geometry

tree







variety

 \mathbb{P}^3

Translate the original problem into toric geometry











 \mathbb{P}^3

Translate the original problem into toric geometry

variety

 \mathbb{P}^3





polytope



projection

Translate the original problem into toric geometry







projection

variety

 \mathbb{P}^3 \mathbb{C}^* action

Translate the original problem into toric geometry





> \prec \sim





projection

variety

 \mathbb{P}^3 \mathbb{C}^* action



Translate the original problem into toric geometry





polytope



projection



variety

 \mathbb{P}^3 \mathbb{C}^* action

Translate the original problem into toric geometry



Translate the original problem into toric geometry





projection



 \mathbb{P}^3 \mathbb{C}^* action

GIT quotient

Translate the original problem into toric geometry


The mutation of a 4-leaf tree



can be explicitly written as deformation which preserves the action of \mathbb{C}^\ast groups associated to leaves,

The mutation of a 4-leaf tree



can be explicitly written as deformation which preserves the action of \mathbb{C}^* groups associated to leaves, thus via GIT quotient it can be extended to a mutation of any tree along any inner edge



a problem to think about (1)

can one see the size a tree by looking at a leaf?

a problem to think about (1)

can one see the size a tree by looking at a leaf?



a problem to think about (2)

a symmetric model of four state system ACTG

a problem to think about (2)

a symmetric model of four state system ACTG with transition matrix

$$\left(\begin{array}{cccc}a&b&c&d\\b&a&d&c\\c&d&a&b\\d&c&b&a\end{array}\right)$$

a problem to think about (2)

a symmetric model of four state system ACTG

0 $0 \ 0 \ 0 \ 0$ $1 \ 0 \ 0$ () $1 \ 0$ ()()1 0 0 1 0 0 () $\left(\right)$ () $\left(\right)$ $\left(\right)$ 0 0 0 1 0 1 0 ()0 $\left(\right)$ ()0 0 1 0 1 0 1 ()()()()1 1 $\left(\right)$ $\left(\right)$ $\left(\right)$ ()() $\left(\right)$ ()()1 1 1 (0 () $\left(\right)$ ()()0 1 0 1 0 0 1 $\left(\right)$ $\left(\right)$ $\mathbf{0}$ $\mathbf{0}$ 1 $\left(\right)$ 0 1 $\left(\right)$ 0 ()0 0 1

a problem to think about (3)

the strand model of four state system ACTG

a problem to think about (3)

the strand model of four state system ACTG with transition matrix

$$\begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}$$