Toric Ideals of Phylogenetic Invariants for the General Group-based Model

Sonja Petrović
(joint work with Julia Chifman)

Mathematics Department
University of Kentucky

ALGEBRAIC BIOLOGY 2007
RISC, Hagenberg

July 4, 2007

Outline

- Motivation:

Outline

- Motivation:

  Statistical models of evolution

Outline

- Motivation:

  Statistical models of evolution

  Algebraic statistics for computational biology

Outline

- Motivation:

  Statistical models of evolution

  Algebraic statistics for computational biology

  Phylogenetic algebraic geometry

Outline

- Motivation:

  Statistical models of evolution

  Algebraic statistics for computational biology

  Phylogenetic algebraic geometry

- Ideals of phylogenetic invariants

Outline

- Motivation:

  Statistical models of evolution

  Algebraic statistics for computational biology

  Phylogenetic algebraic geometry

- Ideals of phylogenetic invariants

- Main Theorem and consequences

Outline

- Motivation:

  Statistical models of evolution

  Algebraic statistics for computational biology

  Phylogenetic algebraic geometry

- Ideals of phylogenetic invariants

- Main Theorem and consequences

- What next...

Motivation

Motivation

Figure: Pictures by Marta Casanellas

Motivation

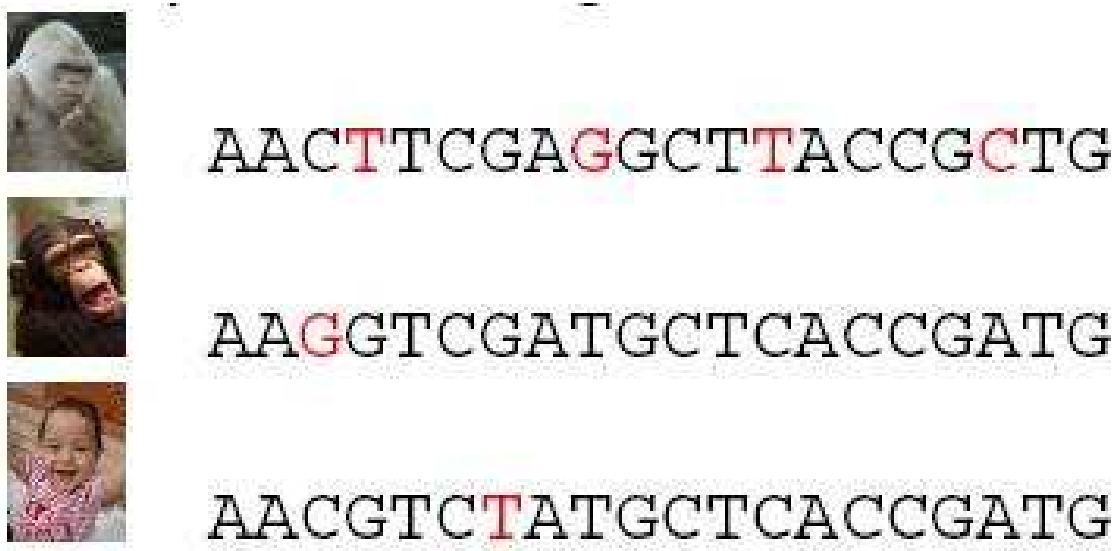


Figure: Pictures by Marta Casanellas

Motivation

Philosophy: statistical model of evolution = algebraic variety

Motivation

Philosophy: statistical model of evolution = algebraic variety

- Statistical inference problem

  DNA sequences $\mapsto$ evolutionary tree

Motivation

Philosophy: statistical model of evolution = algebraic variety

- Statistical inference problem

  DNA sequences $\mapsto$ evolutionary tree

- Probabilistic models on a tree

Motivation

Philosophy: statistical model of evolution = algebraic variety

- Statistical inference problem
  DNA sequences $\mapsto$ evolutionary tree

- Probabilistic models on a tree

- Phylogenetic algebraic geometry picture:
  models $\mapsto$ algebraic varieties

Motivation

Philosophy: statistical model of evolution = algebraic variety

- Statistical inference problem
  DNA sequences $\mapsto$ evolutionary tree
- Probabilistic models on a tree
- Phylogenetic algebraic geometry picture:
  models $\mapsto$ algebraic varieties
- Application
  Use phylogenetic ideals to determine genetic relationship between species based on their DNA sequences.

Statistical models of evolution

GOAL:

construct the phylogenetic tree corresponding to the given species.

Statistical models of evolution

GOAL:

construct the phylogenetic tree corresponding to the given species.

A phylogenetic tree T is a simple, connected, acyclic graph

equipped with some statistical information:

Statistical models of evolution

GOAL:

construct the phylogenetic tree corresponding to the given species.

A phylogenetic tree T is a simple, connected, acyclic graph

equipped with some statistical information:

- each node of T is a random variable with k possible states chosen from the state space S.

Statistical models of evolution

GOAL:

construct the phylogenetic tree corresponding to the given species.

A phylogenetic tree T is a simple, connected, acyclic graph

equipped with some statistical information:

- each node of T is a random variable with k possible states chosen from the state space S.
- Edges of T are labeled by transition probability matrices that reflect probabilities of changes of the states from a node to its child.

Statistical models of evolution

GOAL:

construct the phylogenetic tree corresponding to the given species.

A phylogenetic tree T is a simple, connected, acyclic graph

equipped with some statistical information:

- each node of T is a random variable with k possible states chosen from the state space S.
- Edges of T are labeled by transition probability matrices that reflect probabilities of changes of the states from a node to its child.
- These probabilities of mutation are the parameters for the statistical model of evolution

Statistical models of evolution

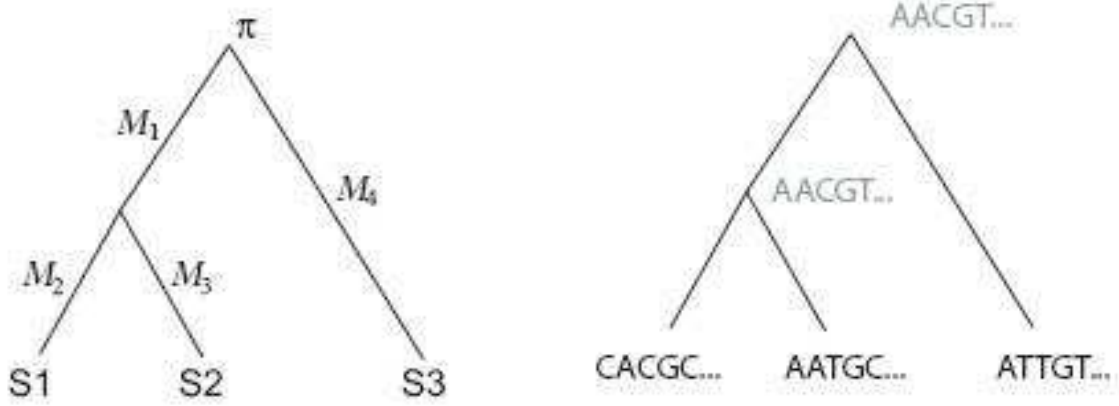
GOAL:

construct the phylogenetic tree corresponding to the given species.

A phylogenetic tree T is a simple, connected, acyclic graph

equipped with some statistical information:

- each node of T is a random variable with k possible states chosen from the state space S.
- Edges of T are labeled by transition probability matrices that reflect probabilities of changes of the states from a node to its child.
- These probabilities of mutation are the parameters for the statistical model of evolution



From DNA to data to joint probabilities

Given n species, determine their genetic relationship.

From DNA to data to joint probabilities

Given n species, determine their genetic relationship.

- DNA sequences $\mapsto$ aligned DNA sequences

Aligned DNA sequences?

- Given DNA sequences, an alignment is a correspondence between them that accounts for their differences.

Aligned DNA sequences?

- Given 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.

Aligned DNA sequences?

- Given 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.
- For example:

Aligned DNA sequences?

- Given 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.
- For example:
- seq1 : ACGTAGCTAAGTTA...
  seq2 : ACCGAGACCCAGTA...
  seq3 : ACCAAGACACAGTA...

Aligned DNA sequences?

- Given 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.
- For example:
- seq1 : ACGTAGCTAAGTTA...
  seq2 : ACCGAGACCCAGTA...
  seq3 : ACCAAGACACAGTA...
- A possible alignment is:
  seq1 : A C – G – T A – G C T A A G T T A
  seq2 : A C C G A G A C – C C A – G T – A
  seq3 : A C C A A G A C A – C A – G T – A...

From DNA to data to joint probabilities

Given n species, determine their genetic relationship.

- DNA sequences $\mapsto$ aligned DNA sequences
- Data = observed pattern frequencies in aligned sequences

From DNA to data to joint probabilities

Given n species, determine their genetic relationship.

- DNA sequences $\mapsto$ aligned DNA sequences
- Data = observed pattern frequencies in aligned sequences

- Example:
  seq1 : A C – G – T A – G C T A A G T T A
  seq2 : A C C G A G A C – C C A – G T – A
  seq3 : A C C A A G A C A – C A – G T – A

$$\hat{p}_{GGA} = \frac{\text{number of observations of GGA}}{\text{sequence length}}$$

The Problem: estimate probability from the data

- (Assume a model of molecular evolution along a tree.)

  p_{GGA} = true probability of observing GGA at a site.

  $\hat{p}$ should estimate the true joint distribution p_{GGA}.

The Problem: estimate probability from the data

- (Assume a model of molecular evolution along a tree.)

  p_{GGA} = true probability of observing GGA at a site.

  $\hat{p}$ should estimate the true joint distribution p_{GGA}.

- IDEA:
  - observe $\hat{p}$
  - conjecture a tree
  - for the tree, compute the true probability p
  - does $\hat{p}$ estimate p?

The Problem: estimate probability from the data

- (Assume a model of molecular evolution along a tree.)

  p_{GGA} = true probability of observing GGA at a site.

  $\hat{p}$ should estimate the true joint distribution p_{GGA}.

- IDEA:
  - observe $\hat{p}$
  - conjecture a tree
  - for the tree, compute the true probability p
  - does $\hat{p}$ estimate p?

- Know the true p for a model and a tree $\Longrightarrow$ plug in data to check if model/tree correct!

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

- focus on 1 site of DNA sequence (assume iid.)

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

- focus on 1 site of DNA sequence (assume iid.)
- assign probabilities to all changes within DNA
  e.g., $p_{A|G}$ = probability of G mutating to an A

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

- focus on 1 site of DNA sequence (assume iid.)
- assign probabilities to all changes within DNA
  e.g., $p_{A|G}$ = probability of G mutating to an A
- nodes of the tree = random variables; values in $S := \{A, G, C, T\}$

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

- focus on 1 site of DNA sequence (assume iid.)
- assign probabilities to all changes within DNA
  e.g., $p_{A|G}$ = probability of G mutating to an A
- nodes of the tree = random variables; values in $S := \{A, G, C, T\}$
- interior nodes = hidden; leaves = observed

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

- focus on 1 site of DNA sequence (assume iid.)
- assign probabilities to all changes within DNA
  e.g., $p_{A|G}$ = probability of G mutating to an A
- nodes of the tree = random variables; values in $S := \{A, G, C, T\}$
- interior nodes = hidden; leaves = observed
- edges of the tree = matrices of probabilities of mutation;
  entries of matrices = unknown parameters

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

- focus on 1 site of DNA sequence (assume iid.)
- assign probabilities to all changes within DNA
  e.g., $p_{A|G}$ = probability of G mutating to an A
- nodes of the tree = random variables; values in $S := \{A, G, C, T\}$
- interior nodes = hidden; leaves = observed
- edges of the tree = matrices of probabilities of mutation;
  entries of matrices = unknown parameters
- relationship between rand. var. encoded by the structure of the tree

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

- focus on 1 site of DNA sequence (assume iid.)
- assign probabilities to all changes within DNA
  e.g., $p_{A|G}$ = probability of G mutating to an A
- nodes of the tree = random variables; values in $S := \{A, G, C, T\}$
- interior nodes = hidden; leaves = observed
- edges of the tree = matrices of probabilities of mutation;
  entries of matrices = unknown parameters
- relationship between rand. var. encoded by the structure of the tree
- p_σ := joint probability of making observation $\sigma \subset S^n$ at the leaves.

Find true probabilities for a given tree

GIVEN: tree T on n leaves, statistical model.
GOAL: describe the joint probabilities of making observations on the leaves.

- focus on 1 site of DNA sequence (assume iid.)
- assign probabilities to all changes within DNA
  e.g., $p_{A|G}$ = probability of G mutating to an A
- nodes of the tree = random variables; values in $S := \{A, G, C, T\}$
- interior nodes = hidden; leaves = observed
- edges of the tree = matrices of probabilities of mutation;
  entries of matrices = unknown parameters
- relationship between rand. var. encoded by the structure of the tree
- p_σ := joint probability of making observation $\sigma \subset S^n$ at the leaves.
- p_σ is a polynomial in the model parameters.

Example: find true probabilities for a given tree

T = tree on 2 leaves.

Suppose root has uniform distribution: $p_{rA} = p_{rG} = p_{rC} = p_{rT} = 1/4$.

- What is p_{AG} = probability of observing A at leaf 1 and G at leaf 2?

Example: find true probabilities for a given tree

T = tree on 2 leaves.
Suppose root has uniform distribution: $p_{rA} = p_{rG} = p_{rC} = p_{rT} = 1/4$.

- What is p_{AG} = probability of observing A at leaf 1 and G at leaf 2?

- Edge transition matrix $M_e = \begin{bmatrix} p_{A|A} & p_{A|G} & p_{A|C} & p_{A|T} \\ p_{G|A} & p_{G|G} & p_{G|C} & p_{G|T} \\ p_{C|A} & p_{C|G} & p_{C|C} & p_{C|T} \\ p_{T|A} & p_{T|G} & p_{T|C} & p_{T|T} \end{bmatrix}$

- $p_{AG} = 1/4 p_{rA} p_{A|A} p_{G|A} + 1/4 p_{rG} p_{A|G} p_{G|G} + 1/4 p_{rC} p_{A|C} p_{G|C} + 1/4 p_T p_{A|T} p_{G|T}$

Example: find true probabilities for a given tree

T = tree on 2 leaves.
Suppose root has uniform distribution: $p_{rA} = p_{rG} = p_{rC} = p_{rT} = 1/4$.

- What is p_{AG} = probability of observing A at leaf 1 and G at leaf 2?

- Edge transition matrix $M_e = \begin{bmatrix} p_{A|A} & p_{A|G} & p_{A|C} & p_{A|T} \\ p_{G|A} & p_{G|G} & p_{G|C} & p_{G|T} \\ p_{C|A} & p_{C|G} & p_{C|C} & p_{C|T} \\ p_{T|A} & p_{T|G} & p_{T|C} & p_{T|T} \end{bmatrix}$

- $p_{AG} = 1/4 p_{rA} p_{A|A} p_{G|A} + 1/4 p_{rG} p_{A|G} p_{G|G} + 1/4 p_{rC} p_{A|C} p_{G|C} + 1/4 p_T p_{A|T} p_{G|T}$

Goal

Understand all the polynomials p_σ for any tree on n leaves.

Phylogenetic invariants

A phylogenetic invariant of the model is a polynomial in the p_σ which vanishes for every choice of model parameters.

Phylogenetic invariants

A phylogenetic invariant of the model is a polynomial in the p_σ which vanishes for every choice of model parameters.

$\{$ invariants $\}$ = prime ideal in $K[p_{\sigma_1}, p_{\sigma_2}, \ldots]$.

Phylogenetic invariants

A phylogenetic invariant of the model is a polynomial in the p_σ which vanishes for every choice of model parameters.

$\{$ invariants $\}$ = prime ideal in $K[p_{\sigma_1}, p_{\sigma_2}, \ldots]$.

Objective: compute this ideal explicitly:

Phylogenetic invariants

A phylogenetic invariant of the model is a polynomial in the p_σ which vanishes for every choice of model parameters.

$\{$ invariants $\}$ = prime ideal in $K[p_{\sigma_1}, p_{\sigma_2}, \dots]$.

Objective: compute this ideal explicitly:

polynomial map $\phi : \mathbb{C}^N \to \mathbb{C}^{k^n}$,

N = total number of model parameters.

Phylogenetic invariants

A phylogenetic invariant of the model is a polynomial in the p_σ which vanishes for every choice of model parameters.

$\{$ invariants $\}$ = prime ideal in $K[p_{\sigma_1}, p_{\sigma_2}, \dots]$.

Objective: compute this ideal explicitly:

polynomial map $\phi : \mathbb{C}^N \to \mathbb{C}^{k^n}$,

N = total number of model parameters.

ϕ depends only on T and k;

its coordinate functions are the k^n polynomials p_σ.

Phylogenetic invariants

A phylogenetic invariant of the model is a polynomial in the p_σ which vanishes for every choice of model parameters.

$\{$ invariants $\}$ = prime ideal in $K[p_{\sigma_1}, p_{\sigma_2}, \dots]$.

Objective: compute this ideal explicitly:

polynomial map $\phi : \mathbb{C}^N \to \mathbb{C}^{k^n}$,

N = total number of model parameters.

ϕ depends only on T and k;

its coordinate functions are the k^n polynomials p_σ.

$\phi \mapsto$ parametrization of an algebraic variety.

Phylogenetic invariants

A phylogenetic invariant of the model is a polynomial in the p_σ which vanishes for every choice of model parameters.

$\{$ invariants $\}$ = prime ideal in $K[p_{\sigma_1}, p_{\sigma_2}, \dots]$.

Objective: compute this ideal explicitly:

polynomial map $\phi : \mathbb{C}^N \to \mathbb{C}^{k^n}$,

N = total number of model parameters.

ϕ depends only on T and k;

its coordinate functions are the k^n polynomials p_σ.

$\phi \mapsto$ parametrization of an algebraic variety.

Definition

Ideal of phylogenetic invariants = $\ker \phi$.

Computing this ideal is hard!

Invariants for group-based models are nice

Is there a class of models for which invariants are particularly nice?

Invariants for group-based models are nice

Is there a class of models for which invariants are particularly nice?

- Let M_e be the $k \times k$ transition probability matrix for edge e of T.

Invariants for group-based models are nice

Is there a class of models for which invariants are particularly nice?

- Let M_e be the $k \times k$ transition probability matrix for edge e of T.
- A group-based model is one in which the matrices M_e are pairwise distinct, but it is required that certain entries coincide.

Invariants for group-based models are nice

Is there a class of models for which invariants are particularly nice?

- Let M_e be the $k \times k$ transition probability matrix for edge e of T.
- A group-based model is one in which the matrices M_e are pairwise distinct, but it is required that certain entries coincide.
- For these models, transition matrices are diagonalizable by the Fourier transform of an abelian group.

Invariants for group-based models are nice

Is there a class of models for which invariants are particularly nice?

- Let M_e be the $k \times k$ transition probability matrix for edge e of T.
- A group-based model is one in which the matrices M_e are pairwise distinct, but it is required that certain entries coincide.
- For these models, transition matrices are diagonalizable by the Fourier transform of an abelian group.
- Examples: Jukes-Cantor, Kimura's one-parameter models used in computational biology.

Invariants for claw trees are enough

Claw tree $T_n := K_{1,n}$ is the complete bipartite graph from one node (the root) to n nodes (the leaves).

Invariants for claw trees are enough

Claw tree $T_n := K_{1,n}$ is the complete bipartite graph from one node (the root) to n nodes (the leaves).

Theorem (Sturmfels-Sullivant)

For group-based models, the ideal of phylogenetic invariants for an arbitrary tree can be computed if the ideals for the claw trees are known.

Invariants for claw trees are enough

Claw tree $T_n := K_{1,n}$ is the complete bipartite graph from one node (the root) to n nodes (the leaves).

Theorem (Sturmfels-Sullivant)

For group-based models, the ideal of phylogenetic invariants for an arbitrary tree can be computed if the ideals for the claw trees are known.

In general, it is an open problem to compute the phylogenetic invariants for a claw tree.

Invariants for claw trees are enough

Claw tree $T_n := K_{1,n}$ is the complete bipartite graph from one node (the root) to n nodes (the leaves).

Theorem (Sturmfels-Sullivant)

For group-based models, the ideal of phylogenetic invariants for an arbitrary tree can be computed if the ideals for the claw trees are known.

In general, it is an open problem to compute the phylogenetic invariants for a claw tree.

We consider the ideal for a general group-based model for the group $\mathbb{Z}_2$.

Invariants for claw trees are enough

Claw tree $T_n := K_{1,n}$ is the complete bipartite graph from one node (the root) to n nodes (the leaves).

Theorem (Sturmfels-Sullivant)

For group-based models, the ideal of phylogenetic invariants for an arbitrary tree can be computed if the ideals for the claw trees are known.

In general, it is an open problem to compute the phylogenetic invariants for a claw tree.

We consider the ideal for a general group-based model for the group $\mathbb{Z}_2$.

Fact

Every group-based model is a specialization of the general group-based model.

Invariants for claw trees, general group-based model

Let q_σ be the image of p_σ under the Fourier transform.

Invariants for claw trees, general group-based model

Let q_σ be the image of p_σ under the Fourier transform.

Definition

the ideal of phylogenetic invariants for the tree T_n is the kernel of the following homomorphism between polynomial rings:

$$\varphi_n : \mathbb{C}[q_{g_1,\ldots,g_n} : g_1,\ldots,g_n \in G] \to \mathbb{C}[a_g^{(i)} : g \in G, i = 1,\ldots,n+1]$$
$$q_{g_1,\ldots,g_n} \mapsto a_{g_1}^{(1)} a_{g_2}^{(2)} \ldots a_{g_n}^{(n)} a_{g_1+g_2+\cdots+g_n}^{(n+1)},$$

where G is a finite group with k elements, each corresponding to a state.

Invariants for claw trees, general group-based model

Let q_σ be the image of p_σ under the Fourier transform.

Definition

the ideal of phylogenetic invariants for the tree T_n is the kernel of the following homomorphism between polynomial rings:

$$\varphi_n : \mathbb{C}[q_{g_1,\ldots,g_n} : g_1,\ldots,g_n \in G] \to \mathbb{C}[a_g^{(i)} : g \in G, i = 1,\ldots,n+1]$$
$$q_{g_1,\ldots,g_n} \mapsto a_{g_1}^{(1)} a_{g_2}^{(2)} \ldots a_{g_n}^{(n)} a_{g_1+g_2+\cdots+g_n}^{(n+1)},$$

where G is a finite group with k elements, each corresponding to a state.

The coordinate $q_{g_1,\ldots,g_n}$ corresponds to observing the element g_1 at the first leaf of T, g_2 at the second, ...

Invariants for claw trees, general group-based model, II

- Phylogenetic invariants form a *toric ideal* in the coordinates q_σ...

Invariants for claw trees, general group-based model, II

- Phylogenetic invariants form a *toric ideal* in the coordinates q_σ...

  ... which can be computed from the corresponding lattice basis ideal by saturation.

Invariants for claw trees, general group-based model, II

- Phylogenetic invariants form a *toric ideal* in the coordinates q_σ...

  ... which can be computed from the corresponding lattice basis ideal by saturation.

- For the group $\mathbb{Z}_2$ on any claw tree, we:

Invariants for claw trees, general group-based model, II

- Phylogenetic invariants form a *toric ideal* in the coordinates q_σ...

  ... which can be computed from the corresponding lattice basis ideal by saturation.

- For the group $\mathbb{Z}_2$ on any claw tree, we:

  describe explicitly the lattice basis ideal

Invariants for claw trees, general group-based model, II

- Phylogenetic invariants form a *toric ideal* in the coordinates q_σ...

  ... which can be computed from the corresponding lattice basis ideal by saturation.

- For the group $\mathbb{Z}_2$ on any claw tree, we:

  describe explicitly the lattice basis ideal

  and a quadratic Gröbner basis of the ideal of invariants (without using saturation).

Example: $n = 3$

$$\varphi : q_{000} \mapsto a_0^{(1)} a_0^{(2)} a_0^{(3)} a_{0+0+0}^{(4)} \qquad q_{100} \mapsto a_1^{(1)} a_0^{(2)} a_0^{(3)} a_{1+0+0}^{(4)}$$

$$q_{001} \mapsto a_0^{(1)} a_0^{(2)} a_1^{(3)} a_{0+0+1}^{(4)} \qquad q_{101} \mapsto a_1^{(1)} a_0^{(2)} a_1^{(3)} a_{1+0+1}^{(4)}$$

$$q_{010} \mapsto a_0^{(1)} a_1^{(2)} a_0^{(3)} a_{0+1+0}^{(4)} \qquad q_{110} \mapsto a_1^{(1)} a_1^{(2)} a_0^{(3)} a_{1+1+0}^{(4)}$$

$$q_{011} \mapsto a_0^{(1)} a_1^{(2)} a_1^{(3)} a_{0+1+1}^{(4)} \qquad q_{111} \mapsto a_1^{(1)} a_1^{(2)} a_1^{(3)} a_{1+1+1}^{(4)}$$

Example: $n = 3$

$$\varphi : q_{000} \mapsto a_0^{(1)} a_0^{(2)} a_0^{(3)} a_{0+0+0}^{(4)} \qquad q_{100} \mapsto a_1^{(1)} a_0^{(2)} a_0^{(3)} a_{1+0+0}^{(4)}$$

$$q_{001} \mapsto a_0^{(1)} a_0^{(2)} a_1^{(3)} a_{0+0+1}^{(4)} \qquad q_{101} \mapsto a_1^{(1)} a_0^{(2)} a_1^{(3)} a_{1+0+1}^{(4)}$$

$$q_{010} \mapsto a_0^{(1)} a_1^{(2)} a_0^{(3)} a_{0+1+0}^{(4)} \qquad q_{110} \mapsto a_1^{(1)} a_1^{(2)} a_0^{(3)} a_{1+1+0}^{(4)}$$

$$q_{011} \mapsto a_0^{(1)} a_1^{(2)} a_1^{(3)} a_{0+1+1}^{(4)} \qquad q_{111} \mapsto a_1^{(1)} a_1^{(2)} a_1^{(3)} a_{1+1+1}^{(4)}$$

φ represented by

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Example: $n = 3$

$\dim \ker(\text{ matrix of } \varphi) = 3$.

The lattice basis is given by the rows of:

$$\begin{bmatrix} 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

Example: $n = 3$

$\dim \ker(\text{ matrix of } \varphi) = 3$.
The lattice basis is given by the rows of:

$$\begin{bmatrix} 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

The corresponding ideal is

$$I_3 = (q_{000}q_{111} - q_{100}q_{011}, q_{001}q_{110} - q_{100}q_{011}, q_{010}q_{101} - q_{100}q_{011}).$$

Example: $n = 3$

$\dim \ker(\text{ matrix of } \varphi) = 3$.
The lattice basis is given by the rows of:

$$\begin{bmatrix} 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

The corresponding ideal is

$$I_3 = (q_{000}q_{111} - q_{100}q_{011}, q_{001}q_{110} - q_{100}q_{011}, q_{010}q_{101} - q_{100}q_{011}).$$

Already saturated, thus equals ideal of invariants.

Example: $n = 3$

$\dim \ker(\text{ matrix of } \varphi) = 3$.
The lattice basis is given by the rows of:

$$\begin{bmatrix} 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

The corresponding ideal is

$$I_3 = (q_{000}q_{111} - q_{100}q_{011}, q_{001}q_{110} - q_{100}q_{011}, q_{010}q_{101} - q_{100}q_{011}).$$

Already saturated, thus equals ideal of invariants.
Let $q_{000} > q_{001} > q_{010} > q_{011} > q_{100} > q_{101} > q_{110} > q_{111}$.

Example: $n = 3$

$\dim \ker(\text{ matrix of } \varphi) = 3$.
The lattice basis is given by the rows of:

$$\begin{bmatrix} 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

The corresponding ideal is

$$I_3 = (q_{000}q_{111} - q_{100}q_{011}, q_{001}q_{110} - q_{100}q_{011}, q_{010}q_{101} - q_{100}q_{011}).$$

Already saturated, thus equals ideal of invariants.
Let $q_{000} > q_{001} > q_{010} > q_{011} > q_{100} > q_{101} > q_{110} > q_{111}$.

Remark

The three generators of I_3 above are a lexicographic Gröbner basis for I_3 , since the initial terms, written with coefficient $+1$ in the above description, are relatively prime so all the S-pairs reduce to zero.

Example: $n = 4$

Ideal of invariants for the 4-leaf claw tree:

$$q_{0000}q_{0111} - q_{0100}q_{0011},\ q_{0001}q_{0110} - q_{0100}q_{0011},\ q_{0010}q_{0101} - q_{0100}q_{0011},$$
$$q_{0000}q_{1011} - q_{1000}q_{0011},\ q_{0001}q_{1010} - q_{1000}q_{0011},\ q_{0010}q_{1001} - q_{1000}q_{0011},$$
$$q_{0000}q_{1101} - q_{1000}q_{0101},\ q_{0001}q_{1100} - q_{1000}q_{0101},\ q_{0100}q_{1001} - q_{1000}q_{0101},$$
$$q_{0000}q_{1110} - q_{1000}q_{0110},\ q_{0010}q_{1100} - q_{1000}q_{0110},\ q_{0100}q_{1010} - q_{1000}q_{0110};$$
$$q_{1000}q_{1111} - q_{1100}q_{1011},\ q_{1001}q_{1110} - q_{1100}q_{1011},\ q_{1010}q_{1101} - q_{1100}q_{1011},$$
$$q_{0100}q_{1111} - q_{1100}q_{0111},\ q_{0101}q_{1110} - q_{1100}q_{0111},\ q_{0110}q_{1101} - q_{1100}q_{0111},$$
$$q_{0010}q_{1111} - q_{1010}q_{0111},\ q_{0011}q_{1110} - q_{1010}q_{0111},\ q_{0110}q_{1011} - q_{1010}q_{0111},$$
$$q_{0001}q_{1111} - q_{1001}q_{0111},\ q_{0011}q_{1101} - q_{1001}q_{0111},\ q_{0101}q_{1011} - q_{1001}q_{0111}.$$
$$q_{0000}q_{1111} - q_{1001}q_{0110},\ q_{0001}q_{1110} - q_{1000}q_{0111},\ q_{0011}q_{1100} - q_{1001}q_{0110},$$
$$q_{0010}q_{1101} - q_{1000}q_{0111},\ q_{0101}q_{1010} - q_{1001}q_{0110},\ q_{0100}q_{1011} - q_{1000}q_{0111}.$$

$2^4 = 16$ variables.
Ideal of invariants: 30 generators.
Lattice basis ideal: 10 generators.

Invariants for an arbitrary claw tree

$\mathcal{G}_n$ is a distinguished set of quadric binomials.

Invariants for an arbitrary claw tree

$\mathcal{G}_n$ is a distinguished set of quadric binomials.

Theorem

For $n \geq 4$, $\quad I_n = (q : q^+ - q^- \in \mathcal{G}_n)$.
In addition, this set of generators can be obtained inductively by lifting the generators corresponding to the various phylogenetic ideals on $n-1$ *leaves.*

Invariants for an arbitrary claw tree

$\mathcal{G}_n$ is a distinguished set of quadric binomials.

Theorem

For $n \geq 4$, $\quad I_n = (q : q^+ - q^- \in \mathcal{G}_n)$.
In addition, this set of generators can be obtained inductively by lifting the generators corresponding to the various phylogenetic ideals on $n-1$ leaves.

Theorem

The set $\mathcal{G}_n$ is a lexicographic Gröbner basis of I_n, for any $n \geq 4$.

Invariants for an arbitrary claw tree

$\mathcal{G}_n$ is a distinguished set of quadric binomials.

Theorem

For $n \geq 4$, $\quad I_n = (q : q^+ - q^- \in \mathcal{G}_n)$.
In addition, this set of generators can be obtained inductively by lifting the generators corresponding to the various phylogenetic ideals on $n-1$ leaves.

Theorem

The set $\mathcal{G}_n$ is a lexicographic Gröbner basis of I_n, for any $n \geq 4$.

Corollary

The coordinate ring of the toric variety whose defining ideal is I_n is Koszul.

Invariants for an arbitrary claw tree

$\mathcal{G}_n$ is a distinguished set of quadric binomials.

Theorem

For $n \geq 4$, $\quad I_n = (q : q^+ - q^- \in \mathcal{G}_n)$.
In addition, this set of generators can be obtained inductively by lifting the generators corresponding to the various phylogenetic ideals on $n-1$ leaves.

Theorem

The set $\mathcal{G}_n$ is a lexicographic Gröbner basis of I_n, for any $n \geq 4$.

Corollary

The coordinate ring of the toric variety whose defining ideal is I_n is Koszul.

Number of variables: 2^n. Number of ideal generators:
$\binom{2^n+1}{2} + \cdots + \binom{2^3+1}{2} - [3^n + \cdots + 3^3] - [\binom{2^{n-1}}{2} + \cdots + \binom{2^2}{2}]$.

Conclusion

- Combined with the main result of Sturmfels and Sullivant, this implies that the phylogenetic ideal of every tree for the group $\mathbb{Z}_2$ has a quadratic Gröbner basis.

Conclusion

- Combined with the main result of Sturmfels and Sullivant, this implies that the phylogenetic ideal of every tree for the group $\mathbb{Z}_2$ has a quadratic Gröbner basis.

- Hence, the coordinate ring of the toric variety is a Koszul algebra.

Conclusion

- Combined with the main result of Sturmfels and Sullivant, this implies that the phylogenetic ideal of every tree for the group $\mathbb{Z}_2$ has a quadratic Gröbner basis.
- Hence, the coordinate ring of the toric variety is a Koszul algebra.
- In addition, the ideals for every tree can be computed explicitly.

Conclusion

- Combined with the main result of Sturmfels and Sullivant, this implies that the phylogenetic ideal of every tree for the group $\mathbb{Z}_2$ has a quadratic Gröbner basis.
- Hence, the coordinate ring of the toric variety is a Koszul algebra.
- In addition, the ideals for every tree can be computed explicitly.
- We are working on extending these results to the group $\mathbb{Z}_2 \times \mathbb{Z}_2$.

...

Tree reconstruction

Fact

Phylogenetic invariants are a powerful tool for tree reconstruction.

Tree reconstruction

Fact

Phylogenetic invariants are a powerful tool for tree reconstruction.

For example, they are used in machine-learning and perform better then many other reconstruction methods.

Tree reconstruction

Fact

Phylogenetic invariants are a powerful tool for tree reconstruction.

For example, they are used in machine-learning and perform better then many other reconstruction methods.

References: Nick Eriksson.

Latent class

Fact

Claw trees with the observation at the root hidden represent latent class models.

Latent class

Fact

Claw trees with the observation at the root hidden represent latent class models.

Fact

Information from the ideal of invariants gives insight into the χ^2 statistic.

Plays a major role in the assessment of the fit and model selection.

Latent class

Fact

Claw trees with the observation at the root hidden represent latent class models.

Fact

Information from the ideal of invariants gives insight into the χ^2 statistic.

Plays a major role in the assessment of the fit and model selection.

References: Steve Feinberg; Yi Zhou.