ALGEBRAIC DISTANCES FOR PHYLOGENETIC TREES (but not only this)

T. González (UIB) M. Llabrés (UIB) J. Rocha (UIB)
F. Rosselló (UIB) G. Valiente (UPC)

Décimo Encuentro de Álgebra Computacional y Aplicaciones, EACA 2006

Sevilla, September 7-9, 2006

Actual topic of this talk

To present an application field for (computer) algebra in computational biology, namely in the comparison of graphs.

The mathematics are so basic that most of this work can be proposed as homework in undergraduate courses (we have done it).

Actual topic of this talk

To present an application field for (computer) algebra in computational biology, namely in the comparison of graphs.

The mathematics are so basic that most of this work can be proposed as homework in undergraduate courses (we have done it).

Graphs in biology

Graphs are ubiquous as models in computational biology:

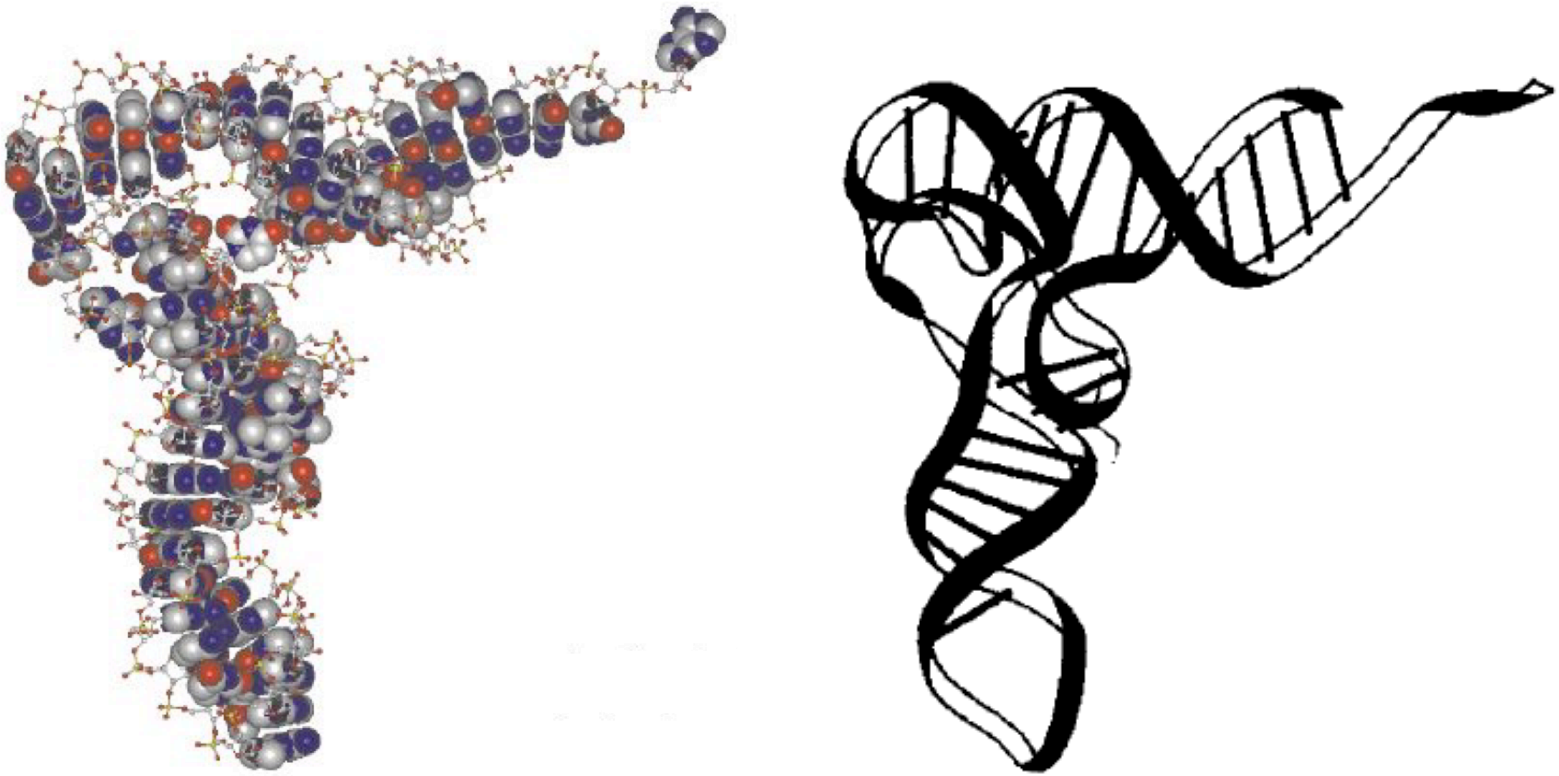
- Models of 3D structures of biopolymers (RNA, proteins)
- Metabolic and other biochemical or genetic networks
- Phylogenetic trees and networks

The (efficient, reliable, meaningful) comparison of these graph models is an important problem in computational biology.

Algebraic models and techniques can be used to define metrics and similarities.

How it all started

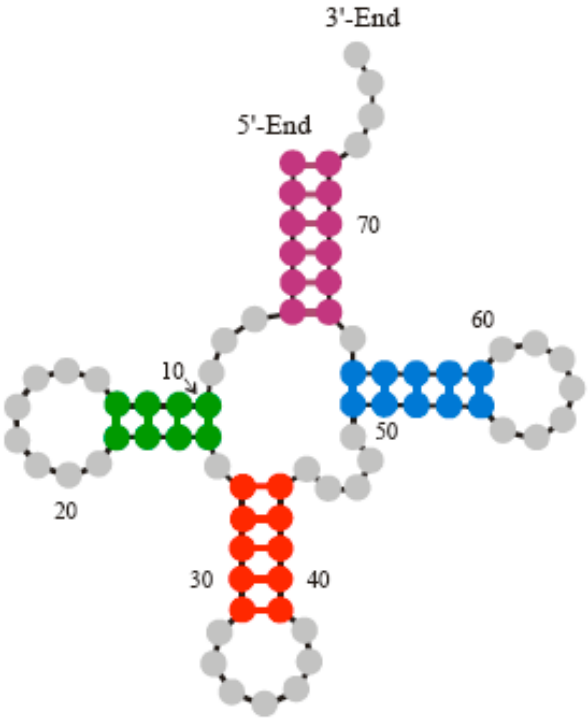
Comparison of RNA 3D structures using algebraic models
(Reidys-Stadler, Comp. & Chem. 20 (1996))

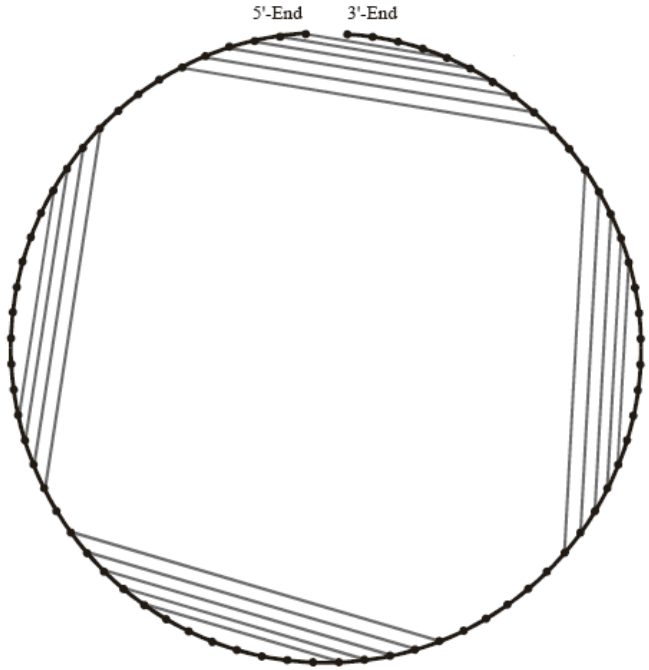


Yeast tRNA$^{\mathrm{Phe}}$ 3D-structure and an abstraction of it

How it all started

Comparison of RNA 3D structures using algebraic models
(Reidys-Stadler, Comp. & Chem. 20 (1996))





Contact structure of yeast $\mathrm{tRNA}^{\mathrm{Phe}}$ 3D-structure

RNA contact structures

An RNA contact structure of length n can be described as an undirected graph $\Gamma = (\{1, \ldots, n\}, B \sqcup Q)$ (B: backbone, Q: contacts) such that:

(i) $B = \big\{\{j, j+1\} \mid j = 1, \ldots, n-1\big\}$

(ii) $\{j, j+1\} \notin Q$ for every j

(iii) Unique bonds: If $\{i, j\}, \{i, k\} \in Q$, then $j = k$.

$\mathcal{C}_n$: the set of all contact structures of length n

$\mathcal{U}_n$: the set of all RNA contact structures of length n.

Reidys-Stadler's work I: Permutation models

An RNA contact structure $\Gamma \in \mathcal{U}_n$ is injectively represented by the permutation

$$\pi(\Gamma) = \prod_{\{i,j\} \in Q} (i,j) \in \mathcal{S}_n$$

Proposition

The mapping $d_{inv} : \mathcal{U}_n \times \mathcal{U}_n \longrightarrow \mathbb{R}$ defined by

$$d_{inv}(\Gamma_1, \Gamma_2) = \textit{least number of transpositions necessary to represent } \pi(\Gamma_2)\pi(\Gamma_1)$$

is a metric on $\mathcal{U}_n$.

Used extensively in Vienna and Santa Fe

Reidys-Stadler's work I: Permutation models

An RNA contact structure $\Gamma \in \mathcal{U}_n$ is injectively represented by the permutation

$$\pi(\Gamma) = \prod_{\{i,j\} \in Q} (i,j) \in \mathcal{S}_n$$

Proposition

The mapping $d_{inv} : \mathcal{U}_n \times \mathcal{U}_n \to \mathbb{R}$ *defined by*

$$d_{inv}(\Gamma_1, \Gamma_2) = \text{least number of transpositions necessary to represent } \pi(\Gamma_2)\pi(\Gamma_1)$$

is a metric on $\mathcal{U}_n$.

Used extensively in Vienna and Santa Fe

Reidys-Stadler's work II: Subgroup models

An RNA contact structure $\Gamma \in \mathcal{U}_n$ is injectively represented by the permutation group

$$G(\Gamma) = \langle (i,j) \mid \{i,j\} \in Q \rangle \subseteq \mathcal{S}_n$$

Proposition

The mapping $d_{sgr} : \mathcal{U}_n \times \mathcal{U}_n \to \mathbb{R}$ defined by

$$d_{sgr}(\Gamma_1, \Gamma_2) = \log_2 \left| \frac{G(\Gamma_1) \cdot G(\Gamma_2)}{G(\Gamma_1) \cap G(\Gamma_2)} \right|$$

is a metric on $\mathcal{U}_n$.

Not too interesting . . . as $d_{sgr}(\Gamma_1, \Gamma_2) = |Q_1 \Delta Q_2|$.

Reidys-Stadler's work II: Subgroup models

An RNA contact structure $\Gamma \in \mathcal{U}_n$ is injectively represented by the permutation group

$$G(\Gamma) = \langle (i,j) \mid \{i,j\} \in Q \rangle \subseteq \mathcal{S}_n$$

Proposition

The mapping $d_{sgr} : \mathcal{U}_n \times \mathcal{U}_n \to \mathbb{R}$ *defined by*

$$d_{sgr}(\Gamma_1, \Gamma_2) = \log_2 \left| \frac{G(\Gamma_1) \cdot G(\Gamma_2)}{G(\Gamma_1) \cap G(\Gamma_2)} \right|$$

is a metric on $\mathcal{U}_n$.

Not too interesting ... as $d_{sgr}(\Gamma_1, \Gamma_2) = |Q_1 \Delta Q_2|$.

Reidys-Stadler's work II: Subgroup models

An RNA contact structure $\Gamma \in \mathcal{U}_n$ is injectively represented by the permutation group

$$G(\Gamma) = \langle (i,j) \mid \{i,j\} \in Q \rangle \subseteq \mathcal{S}_n$$

Proposition

The mapping $d_{sgr} : \mathcal{U}_n \times \mathcal{U}_n \to \mathbb{R}$ *defined by*

$$d_{sgr}(\Gamma_1, \Gamma_2) = \log_2 \left| \frac{G(\Gamma_1) \cdot G(\Gamma_2)}{G(\Gamma_1) \cap G(\Gamma_2)} \right|$$

is a metric on $\mathcal{U}_n$.

Not too interesting ... as $d_{sgr}(\Gamma_1, \Gamma_2) = |Q_1 \Delta Q_2|$.

A monomial distance

To get rid of the unique bonds condition, and to obtain non-trivial metrics, we move from subgroups to monomial ideals (Llabrés-Rosselló, Comp. Biol. Chem. 28 (2004))

A contact structure $\Gamma \in \mathcal{C}_n$ is injectively represented by the edge ideal

$$I_\Gamma = \langle x_i x_j \mid \{i,j\} \in Q \rangle \subseteq \mathbb{F}_2[x_1, \ldots, x_n]$$

Some notations:

π_m: projection modulo all weight m monomials

$M(I)_k$: monomials of weight $\leqslant k$ in I

$H(I)$: Hilbert polynomial of I

A monomial distance

To get rid of the unique bonds condition, and to obtain non-trivial metrics, we move from subgroups to monomial ideals (Llabrés-Rosselló, Comp. Biol. Chem. 28 (2004))

A contact structure $\Gamma \in \mathcal{C}_n$ is injectively represented by the edge ideal

$$I_\Gamma = \langle x_i x_j \mid \{i,j\} \in Q \rangle \subseteq \mathbb{F}_2[x_1, \ldots, x_n]$$

Some notations:

π_m: projection modulo all weight m monomials

$M(I)_k$: monomials of weight $\leqslant k$ in I

$H(I)$: Hilbert polynomial of I

A monomial distance

Proposition

For every $m \geqslant 3$, the mapping $D_m : \mathcal{C}_n \times \mathcal{C}_n \longrightarrow \mathbb{R}$ defined by

$$\begin{aligned} D_m(\Gamma_1, \Gamma_2) &= \log_2 \left| \frac{\pi_m(I_{\Gamma_1}) + \pi_m(I_{\Gamma_2})}{\pi_m(I_{\Gamma_1}) \cap \pi_m(I_{\Gamma_2})} \right| \qquad (D_3 = d_{sgr}) \\ &= |M(I_{\Gamma_1})_{m-1} \Delta M(I_{\Gamma_1})_{m-1}| \\ &= H_{I_{\Gamma_1}}(m-1) + H_{I_{\Gamma_2}}(m-1) - 2H_{I_{\Gamma_1}+I_{\Gamma_2}}(m-1). \end{aligned}$$

is a metric on $\mathcal{C}_n$.

See the survey (Rosselló, in *Recent results in natural computing* (Ed. Fénix, 2005)) for details, other metrics, open problems, etc.

A monomial distance

Proposition

For every $m \geqslant 3$, the mapping $D_m : \mathcal{C}_n \times \mathcal{C}_n \longrightarrow \mathbb{R}$ defined by

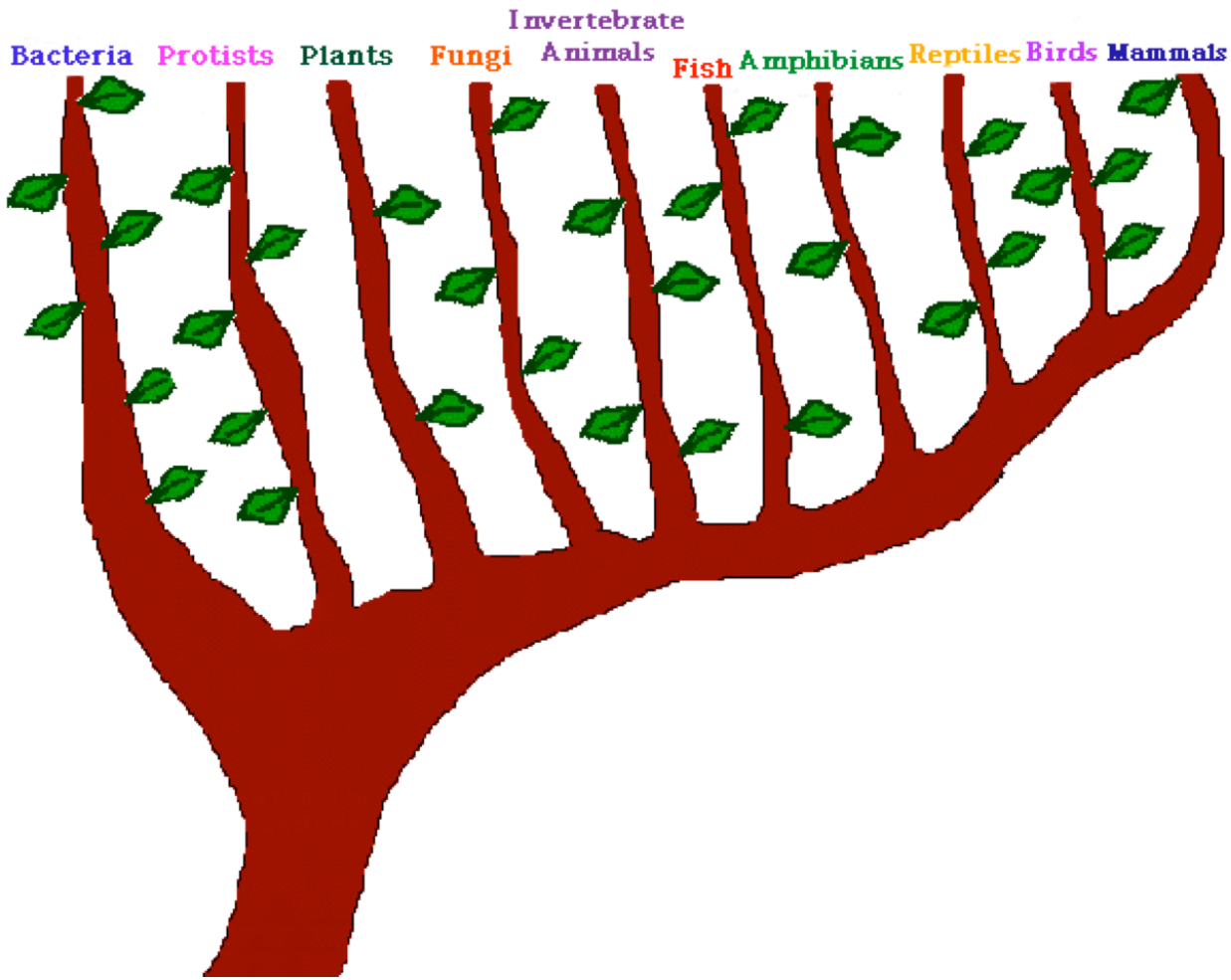
$$\begin{aligned} D_m(\Gamma_1, \Gamma_2) &= \log_2 \left| \frac{\pi_m(I_{\Gamma_1}) + \pi_m(I_{\Gamma_2})}{\pi_m(I_{\Gamma_1}) \cap \pi_m(I_{\Gamma_2})} \right| \qquad (D_3 = d_{sgr}) \\ &= |M(I_{\Gamma_1})_{m-1} \Delta M(I_{\Gamma_1})_{m-1}| \\ &= H_{I_{\Gamma_1}}(m-1) + H_{I_{\Gamma_2}}(m-1) - 2H_{I_{\Gamma_1}+I_{\Gamma_2}}(m-1). \end{aligned}$$

is a metric on $\mathcal{C}_n$.

See the survey (Rosselló, in *Recent results in natural computing* (Ed. Fénix, 2005)) for details, other metrics, open problems, etc.

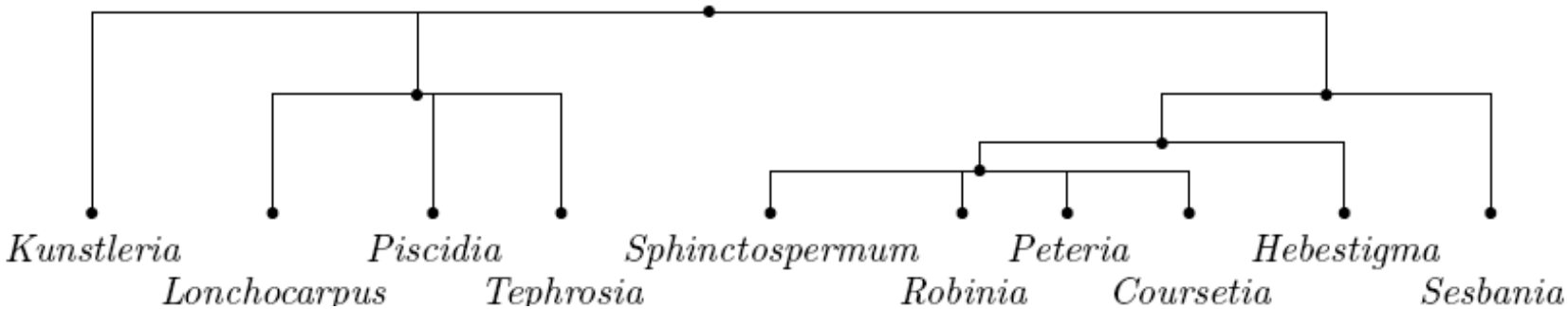
Phylogenetic trees

A phylogenetic tree is a representation of the evolutive dependence and branching of the organisms represented by the leaves.



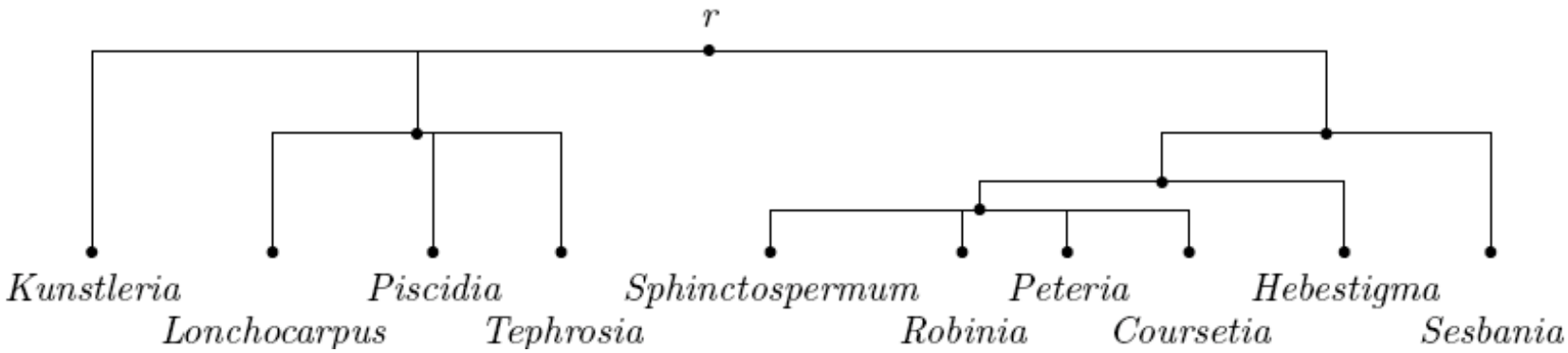
Phylogenetic trees

A phylognetic tree is a *rooted tree with injectively labeled leaves and without outdegree 1 nodes*.



Phylogenetic trees

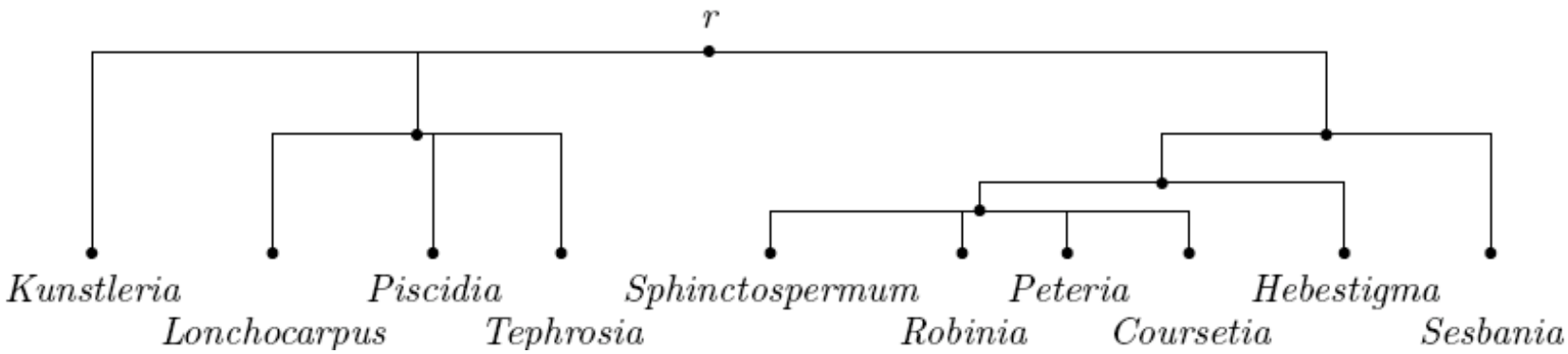
A phylogenetic tree is a *rooted tree with injectively labeled leaves and without outdegree 1 nodes*.



A phylogenetic tree is a directed finite graph $T = (V, E)$ containing a distinguished node $r \in V$, the root, such that for every other node $v \in V$ there exists one, and only one, path from the root r to v.

Phylogenetic trees

A phylogenetic tree is a *rooted tree with injectively labeled leaves and without outdegree 1 nodes*.

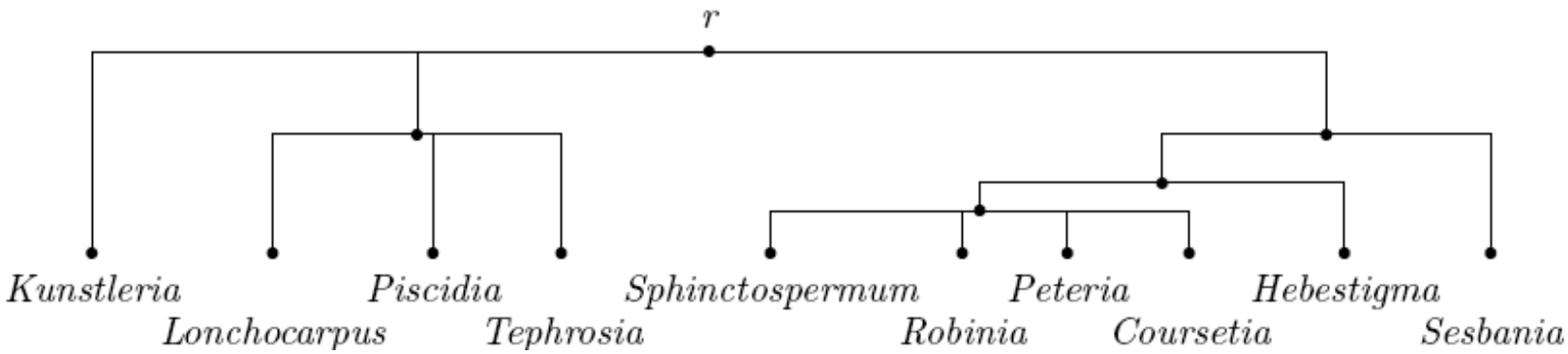


The children of a node v are those nodes $w \in V$ such that $(v, w) \in E$.

The descendants of a node v are those nodes $w \in V$ that can be reached from v through a (directed) path

Phylogenetic trees

A phylogenetic tree is a *rooted tree with injectively labeled leaves and without outdegree 1 nodes*.

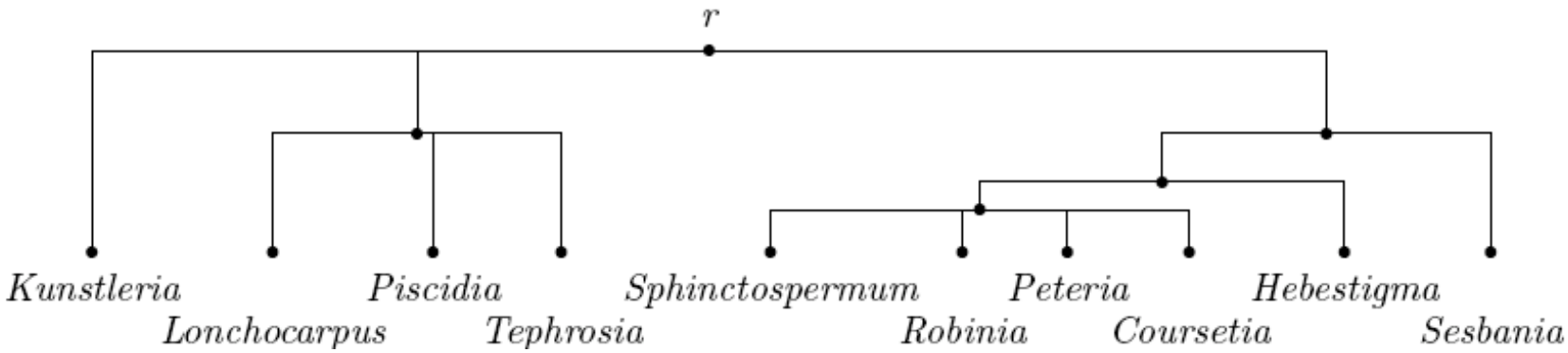


The nodes without children are the leaves of the tree. The set of leaves of T is denoted by $\mathcal{L}(T)$.

The nodes that are not leaves are called internal. We assume that every internal node has at least 2 children.

Phylogenetic trees

A phylogenetic tree is a *rooted tree with injectively labeled leaves and without outdegree 1 nodes*.



The height of a node v is the length of a longest directed path from v to a leaf.

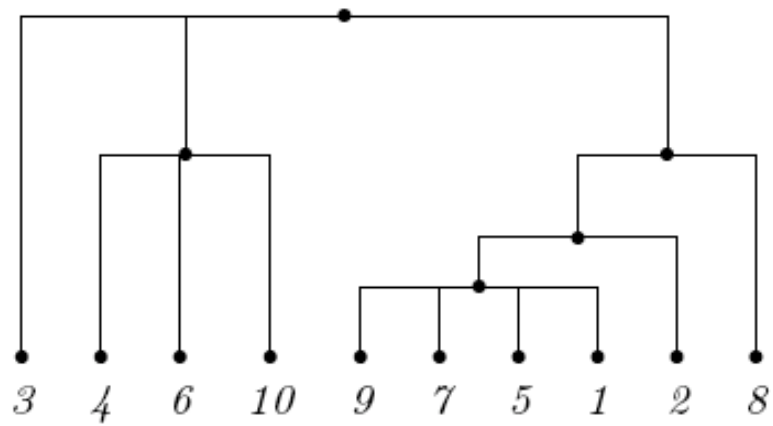
The depth of a node v is the length of the path from r to it.

Phylogenetic trees

The leaves of a phylogenetic tree are injectively labeled in a fixed, but arbitrary, set. In practice, if the tree has n leaves, we shall identify their labels with $1, \ldots, n$.

The label associated to a leaf $v \in V$ will be denoted by $\ell(v)$.

$\mathcal{T}_n$: the set of all phylogenetic trees with n leaves labeled $1, \ldots, n$



Phylogenetic trees: Transposition distance

We assume the leaves ordered $1 < \cdots < n$.
The bottom-up ordering of $T = (V, E) \in \mathcal{T}_n$ is the injective mapping

$$\ell_T : V \to \{1, \ldots, |V|\}$$

defined by:

(a) If $v \in \mathcal{L}(T)$, then $\ell_T(v)$ is its label

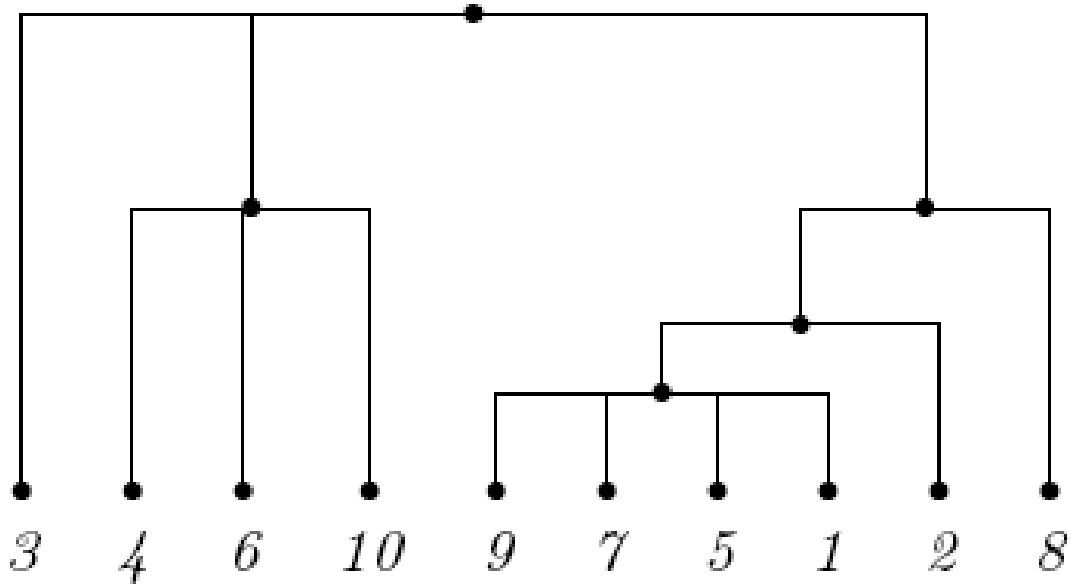
(b) If $height(u) < height(v)$, then $\ell_T(u) < \ell_T(v)$;

(c) If $0 < height(u) = height(v)$ and

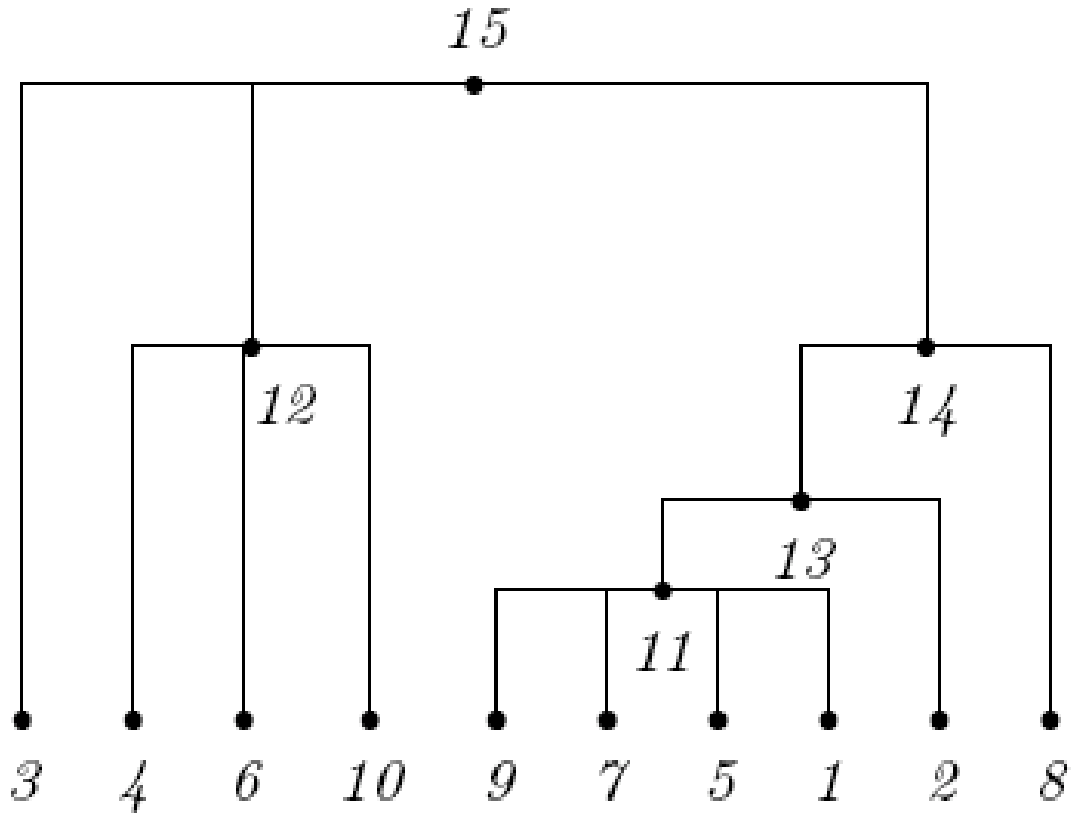
$$\min\{\ell_T(x) \mid x \in children(u)\} < \min\{\ell_T(x) \mid x \in children(v)\},$$

then $\ell_T(u) < \ell_T(v)$.

The previous phylogenetic tree



The previous phylogenetic tree and its bottom-up ordering

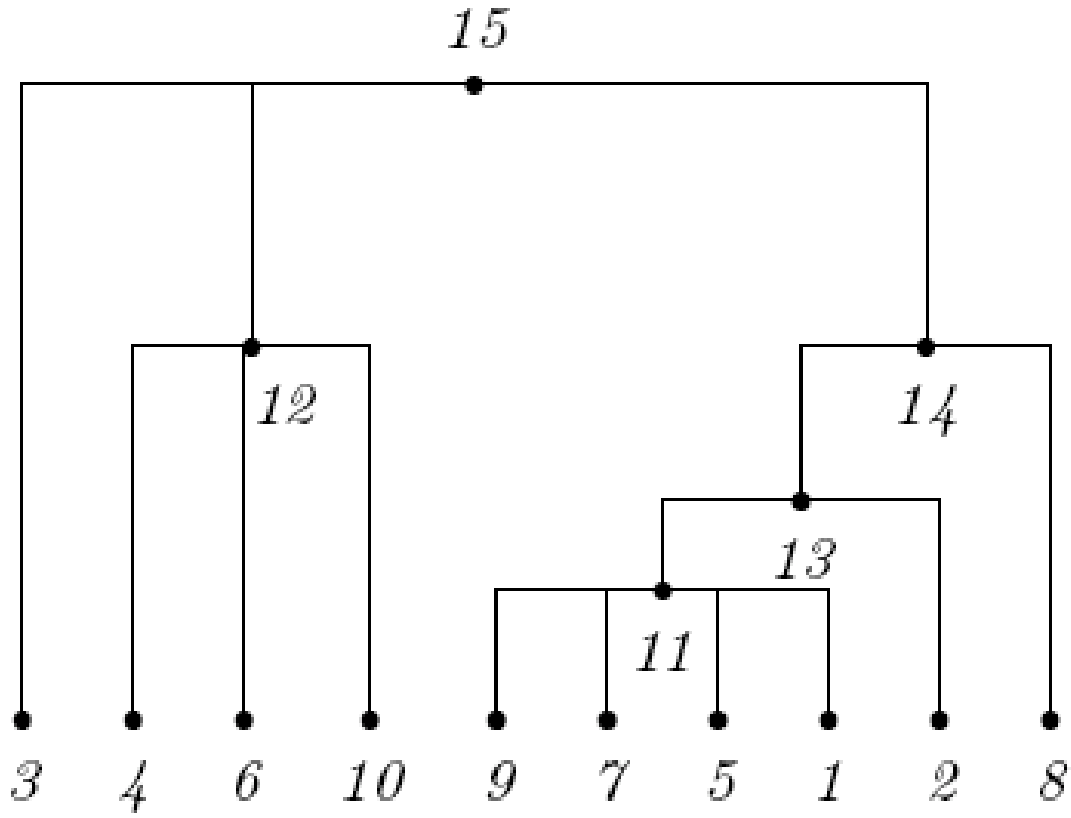


Phylogenetic trees: Transposition distance

The matching representation of a phylogenetic tree $T = (V, E) \in \mathcal{T}_n$ is the partition of $\{1, \ldots, |V| - 1\}$

$$M(T) = \{\ell_T(\mathit{children}(u)) \mid u \in V - \mathcal{L}(T)\}.$$

The previous phylogenetic tree and its matching representation



$$M(T) = \Big\{\{4,6,10\},\{9,7,5,1\},\{11,2\},\{13,8\},\{3,12,14\}\Big\}$$

Phylogenetic trees: Transposition distance

The matching representation of a phylogenetic tree $T = (V, E) \in \mathcal{T}_n$ is the partition of $\{1, \ldots, |V| - 1\}$

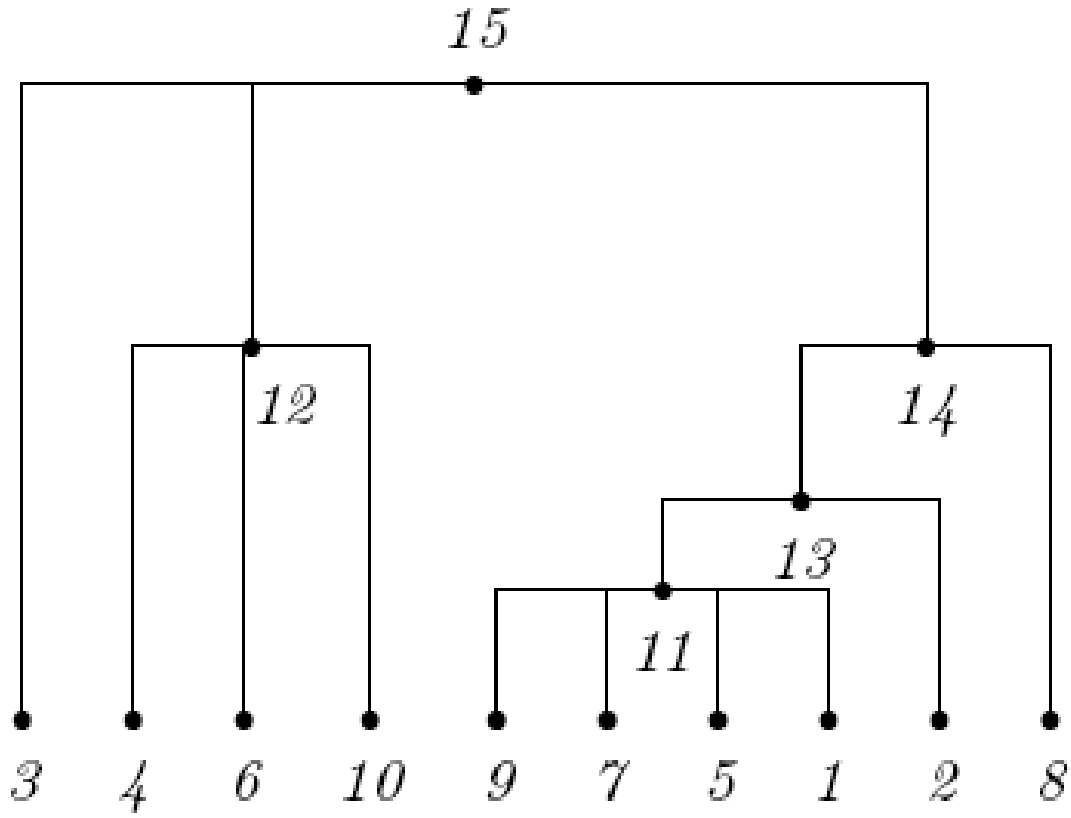
$$M(T) = \{\ell_T(children(u)) \mid u \in V - \mathcal{L}(T)\}.$$

The cycle associated to an ordered set $S = \{i_1, \ldots, i_k\}$, with $i_1 < \cdots < i_k$ and $k \geqslant 2$, is $\kappa(S) := (i_1, i_2, \ldots, i_k)$.

The matching permutation $\pi(T)$ associated to a phylogenetic tree $T = (V, E) \in \mathcal{T}_n$ is the permutation of $\{1, \ldots, 2n - 2\}$ defined by the product of the cycles associated to the members of its matching representation:

$$\pi(T) = \prod_{u \in V - \mathcal{L}(T)} \kappa(\ell_T(children(u))) \in \mathcal{S}_{2n-2}.$$

The previous phylogenetic tree and its matching permutation



$\pi(T) = (4, 6, 10)(1, 5, 7, 9)(2, 11)(8, 13)(3, 12, 14)$

Phylogenetic trees: Transposition distance

Proposition

The mapping $TD : \mathcal{T}_n \times \mathcal{T}_n \to \mathbb{R}$ *defined by*

$$TD(T_1, T_2) = \text{least number of transpositions necessary to represent } \pi(T_2)^{-1}\pi(T_1)$$

is a metric on $\mathcal{T}_n$.

It takes values $0, 2, 4, 6, \ldots, 2n-4$, and it can be computed in linear time

More details in Rosselló, Valiente, q-bio/0604024

Main drawback: it depends too much on the label ordering

Phylogenetic trees: Monomial distances

To get rid of the label ordering, we use suitable monomial ideals. For instance:

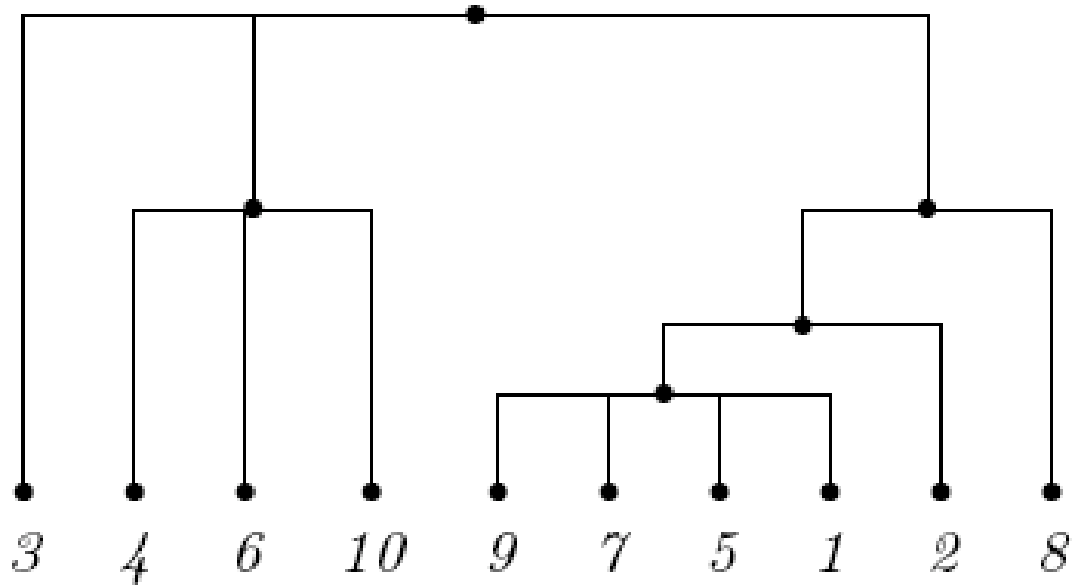
The monomial associated to $S = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ is $\mu(S) := x_{i_1} \cdots x_{i_k}$.

The monomial associated to an internal node v of $T \in \mathcal{T}_n$ is $\mu(v) := x_0^{\mathrm{depth}(v)} \mu(\mathcal{L}(v))$, where $\mathcal{L}(v)$ stands for the set of labels of the leaves that are descendants of v.

The monomial ideal I_T associated to $T = (V, E) \in \mathcal{T}_n$ is

$$I_T = \langle \mu(v) \mid v \in V - (\mathcal{L}(T) \cup \{r\}) \rangle \subseteq \mathbb{F}_2[x_0, x_1, \ldots, x_n]$$

The previous phylogenetic tree and its monomial ideal



$I_T = \langle x_0x_4x_6x_{10}, x_0^3x_1x_5x_7x_9, x_0^2x_1x_2x_5x_7x_9, x_0x_1x_2x_5x_7x_8x_9 \rangle$

(without x_0, we loose information)

Proposition

The mapping $MD_n : \mathcal{T}_n \times \mathcal{T}_n \to \mathbb{N}$ *defined by*

$$MD_n(T_1, T_2) = |M(I_{T_1})_n \Delta M(I_{T_2})_n|$$

is a metric on $\mathcal{T}_n$.

Too large numbers, quite expensive to compute, but it refines other edit distances.

One could "count" instead monomials of weight n square-free in $x_1, \ldots, x_n$

For more details and other monomial distances, browse the arXiv in a near future

Open problems

- Use of other algebraic objects
- Efficient computation of monomial distances
- Comparison of contact structures of different lengths
- Comparison of phylogenetic trees with nested taxa (labeled internal nodes)
- Comparison of phylogenetic networks (completely unexplored)
- ...