DCJ-indel distance of natural genomes

Leonard Bohnenkämper, Marília D. V. Braga

January 28, 2021

Types of genomes

Given a pair of genomes $\mathbb{A}, \mathbb{B}$. Let $\Phi_G(m)$ be the copy number of family m in genome $G \in \{\mathbb{A}, \mathbb{B}\}$.

	Non-Singular $\Phi_G(m)$ arbitrary	Singular $\Phi_G(m) \leq 1$
Unbalanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert$ arbitrary	Natural genomes	Singular Genomes
Balanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert = 0$	Balanced Genomes	Canonical Genomes

Types of genomes

Given a pair of genomes $\mathbb{A}, \mathbb{B}$. Let $\Phi_G(m)$ be the copy number of family m in genome $G \in \{\mathbb{A}, \mathbb{B}\}$.

	Non-Singular $\Phi_G(m)$ arbitrary	Singular $\Phi_G(m) \leq 1$
Unbalanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert$ arbitrary		
Balanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert = 0$		

Types of genomes

Given a pair of genomes $\mathbb{A}, \mathbb{B}$. Let $\Phi_G(m)$ be the copy number of family m in genome $G \in \{\mathbb{A}, \mathbb{B}\}$.

	Non-Singular $\Phi_G(m)$ arbitrary	Singular $\Phi_G(m) \leq 1$
Unbalanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert$ arbitrary		
Balanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert = 0$		original DCJ model

Types of genomes

Given a pair of genomes $\mathbb{A}, \mathbb{B}$. Let $\Phi_G(m)$ be the copy number of family m in genome $G \in \{\mathbb{A}, \mathbb{B}\}$.

	Non-Singular $\Phi_G(m)$ arbitrary	Singular $\Phi_G(m) \leq 1$
Unbalanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert$ arbitrary		DCJ-Indel model
Balanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert = 0$		original DCJ model

Types of genomes

Given a pair of genomes $\mathbb{A}, \mathbb{B}$. Let $\Phi_G(m)$ be the copy number of family m in genome $G \in \{\mathbb{A}, \mathbb{B}\}$.

	Non-Singular $\Phi_G(m)$ arbitrary	Singular $\Phi_G(m) \leq 1$
Unbalanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert$ arbitrary		DCJ-Indel model
Balanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert = 0$	ILP by Shao et al.	original DCJ model

Types of genomes

Given a pair of genomes $\mathbb{A}, \mathbb{B}$. Let $\Phi_G(m)$ be the copy number of family m in genome $G \in \{\mathbb{A}, \mathbb{B}\}$.

	Non-Singular $\Phi_G(m)$ arbitrary	Singular $\Phi_G(m) \leq 1$
Unbalanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert$ arbitrary	ILP today	DCJ-Indel model
Balanced $\lvert\Phi_{\mathbb{A}}(m) - \Phi_{\mathbb{B}}(m)\rvert = 0$	ILP by Shao et al.	original DCJ model

Types of genomes

Given a pair of genomes A, B. Let $\Phi_G(m)$ be the copy number of family m in genome $G \in \{A, B\}$.

	Non-Singular $\Phi_G(m)$ arbitrary	Singular $\Phi_G(m) \leq 1$
Unbalanced $\lvert\Phi_A(m) - \Phi_B(m)\rvert$ arbitrary	ILP today	DCJ-Indel model
Balanced $\lvert\Phi_A(m) - \Phi_B(m)\rvert = 0$	ILP by Shao et al.	original DCJ model

NP-Hard

Natural Genomes and indels

There is no way to sort ($1\ 2\ 2\ \bar{3}\ 4$) into ($1\ 2\ 3\ 2\ 2\ 4$) by DCJs alone.

Natural Genomes and indels

There is no way to sort ($1\ 2\ 2\ \bar{3}\ 4$) into ($1\ 2\ 3\ 2\ 2\ 4$) by DCJs alone.

We need indel operations

Natural Genomes and indels

There is no way to sort ($1\ 2\ 2\ \bar{3}\ 4$) into ($1\ 2\ 3\ 2\ 2\ 4$) by DCJs alone.

We need indel operations

$\rightarrow$ But how many 2s to delete/insert?

Excursus: Matching models

This depends on the assumed phylogeny!

$$\begin{tikzcd} & ? \arrow[dl] \arrow[dr] & \\ (1\ 2\ 2\ \bar{3}\ 4) & & (1\ 2\ 3\ 2\ 2\ 4) \end{tikzcd}$$

Excursus: Matching models

This depends on the assumed phylogeny!

$$\begin{tikzcd} & (1\ 2\ 3\ 4) \arrow[dl] \arrow[dr] & \\ (1\ 2\ 2\ \bar{3}\ 4) & & (1\ 2\ 3\ 2\ 2\ 4) \end{tikzcd}$$

Excursus: Matching models

This depends on the assumed phylogeny!

$$\begin{tikzcd} & (1\ 2\ 3\ \bar{2}\ 4) \arrow[dl] \arrow[dr] & \\ (1\ 2\ 2\ \bar{3}\ 4) & & (1\ 2\ 3\ 2\ 2\ 4) \end{tikzcd}$$

Excursus: Matching models

This depends on the assumed phylogeny!

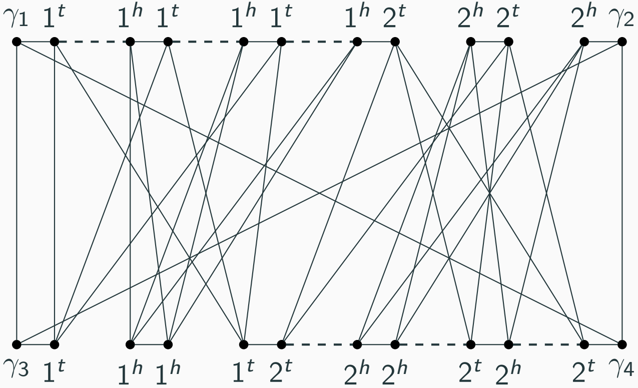
$(1\ 2\ 2\ \bar{3}\ 4)$ $(1\ 2\ 3\ 2\ 2\ 4)$

Excursus: Matching models

How to handle a shared family m with $\Phi_A(m) \neq \Phi_B(m)$

Exemplary Matching (EM)	Intermediate Matching (IM)	Maximal Matching (MM)
Exactly one occurrence matched	At least one occurrence matched	As many occurrences as possible matched
$n_m = 1$ genes of family m matched	$1 \leq n_m \leq min(\Phi_A(m), \Phi_B(m))$ genes of family m matched	$n_m = min(\Phi_A(m), \Phi_B(m))$ genes of family m matched
Lowest common ancestor: Each shared marker occurs once		Lowest common ancestor: Each shared marker occurs at least as often as in the genome with fewer occurrences

Enforcing MM in the capped MRG



The capped MRG for MM Natural Genomes

Given two natural genomes $\mathbb{A}$, $\mathbb{B}$ their *capped multi-relational graph* $CMRG(\mathbb{A}, \mathbb{B})$ is described as follows

1. $V = V(\xi(\mathbb{A})) \cup V(\xi(\mathbb{B})) \cup \Gamma$: There is a vertex for each extremity/cap in each genome.
   Each vertex v has a label $\ell(v)$ corresponding to the extremity it represents.
2. $E = E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B}) \cup E_\xi \cup E_{\xi'} \cup E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
   - $E_\alpha(\mathbb{G}) = \{uv : u, v \in V(\xi(\mathbb{G}))$ and $\ell(u)\ell(v) \in \alpha(\mathbb{G})\}$
   - $E_\xi = \{uv : u \in V(\xi(\mathbb{A}))$ and $v \in V(\xi(\mathbb{B}))$ and $\ell(u) = \ell(v)\}$
   - $E_{\xi'}$ ... edges connecting caps
   -

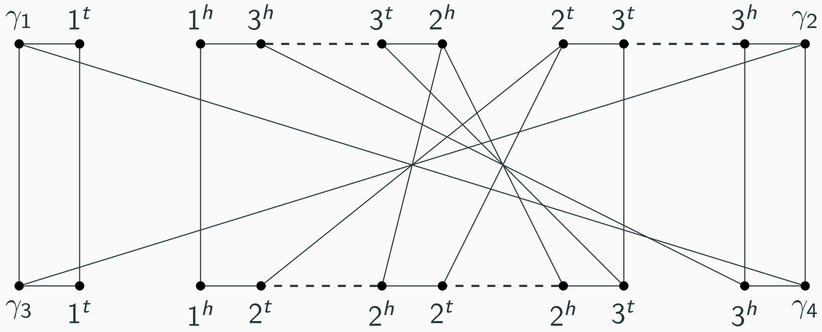
$$E_{ID}(\mathbb{G}) = \{uv : u, v \in V(\xi(\mathbb{G})) \text{ and } u, v \text{ are extremities of the same gene of family } m \text{ with } \Phi_{\mathbb{G}}(m) > min(\Phi_{\mathbb{A}}(m), \Phi_{\mathbb{B}}(m))\}$$

Consistent decompositions in the CMRG

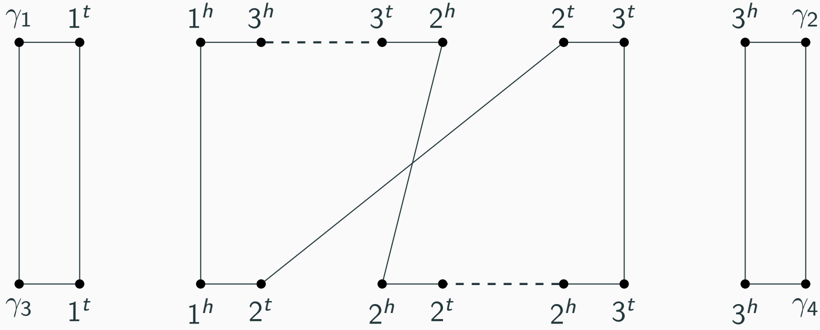
Capped consistent decomposition $Q[S,P]$

- is **induced** by a maximal sibling-set S and a maximal capping-set P
- is the union of S with P with all adjacency edges and indel edges of genes not matched in S
- covers all vertices of $CMRG(\mathbb{A},\mathbb{B})$
- is composed of cycles only

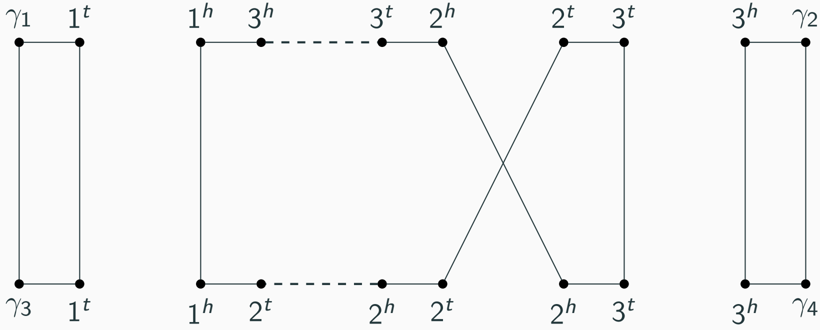
Consistent Decompositions $\equiv$ Matchings



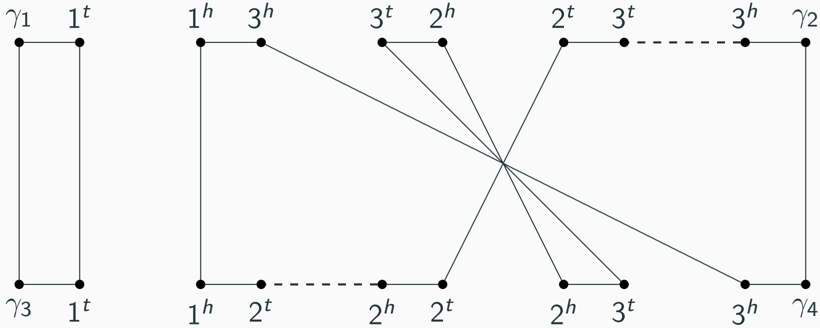
Consistent Decompositions $\equiv$ Matchings



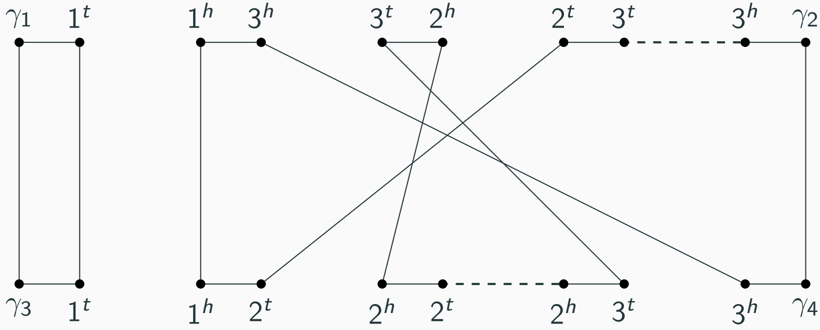
Consistent Decompositions $\equiv$ Matchings



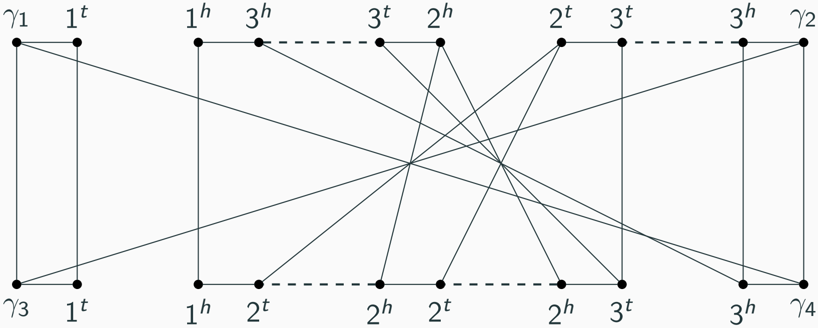
Consistent Decompositions $\equiv$ Matchings



Consistent Decompositions $\equiv$ Matchings



Consistent Decompositions $\equiv$ Matchings



Finding the best decomposition

$$d_{DCJ}^{ID}(A,B) = \min_{S \in \mathfrak{S}_{\mathrm{MAX}}, P \in \mathfrak{P}_{\mathrm{MAX}}} \{d_{DCJ}^{ID}(Q[S,P])\} = n_* + p_* - \max_{S \in \mathfrak{S}_{\mathrm{MAX}}, P \in \mathfrak{P}_{\mathrm{MAX}}} \{w(Q[S,P])\}\ ,$$

$$\text{where} \begin{cases} \mathfrak{S}_{\mathrm{MAX}} \text{ is the set of all maximal sibling-sets of } CMRG(\mathbb{A},\mathbb{B}) \\ \mathfrak{P}_{\mathrm{MAX}} \text{ is the set of all maximal capping-sets of } CMRG(\mathbb{A},\mathbb{B}) \\ n_* \text{ and } p_* \text{ are constant for any capped consistent decomposition} \end{cases}$$

with $w(Q[S,P]) = |\mathcal{C}^Q| - \sum_{C \in \mathcal{C}^Q \cup \mathcal{S}^Q} (\lambda(C))$

where

$\mathcal{C}^Q$ are cycles containing extremity edges

$\mathcal{S}^Q$ are circular singletons

Recap: Shao-Lin-Moret

Match the parts of the ILP to their function!

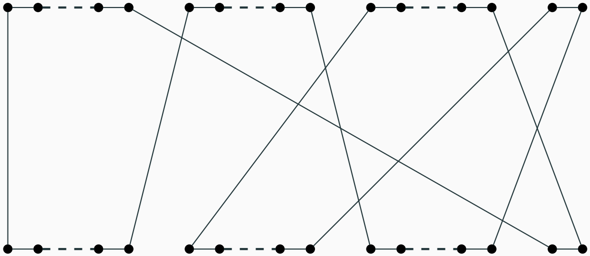
A	$\ell_i \leq \ell_j + i(1 - x_{\{v_i, v_j\}})$	$\forall \{v_i, v_j\} \in E$
B	$\sum_{\{u,v\} \in E} x_{\{u,v\}} = 2$	$\forall u \in V$
C	$i \cdot z_i \leq \ell_i$	$\forall 1 \leq i \leq \lvert V \rvert$
D	$x_e = 1$	$\forall e \in E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$
E	$x_e = x_d$	$\forall e, d \in E_\xi$ such that e and d are siblings

1. Each adjacency edge is in the decomposition
2. Sibling edges are only selected together
3. A cycle is only counted at the vertex with the smalles label
4. A decomposition consists only of simple cycles
5. Cycle labels of adjacent vertices are the same

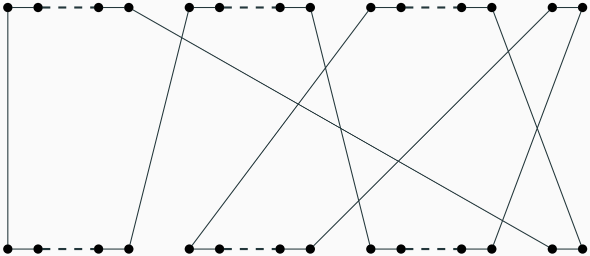
Recap: Capping and indels

	id	sources	linking AB-cycle	T	$\boldsymbol{\Delta n}$	$\boldsymbol{\Delta c}$	$\boldsymbol{\Delta (2i)}$	$\boldsymbol{\Delta\lambda}$	$\boldsymbol{\Delta d}$
$\mathcal{P}$	$\mathtt{WM}$	$AA_{\mathcal{AB}} + BB_{\mathcal{AB}}$	$(AA_{\mathcal{AB}}, BB_{\mathcal{BA}})$		$+1$	$+1$	0	-2	-2
$\mathcal{Q}$	$\mathtt{WW\overline{M}\underline{M}}$	$2 \times AA_{\mathcal{AB}} + BB_{\mathcal{A}} + BB_{\mathcal{B}}$	$(AA_{\mathcal{AB}}, BB_{\mathcal{B}}, AA_{\mathcal{BA}}, BB_{\mathcal{A}})$		$+2$	$+1$	0	-4	-3
	$\mathtt{MM\overline{W}\underline{W}}$	$2 \times BB_{\mathcal{AB}} + AA_{\mathcal{A}} + AA_{\mathcal{B}}$	$(BB_{\mathcal{AB}}, AA_{\mathcal{B}}, BB_{\mathcal{BA}}, AA_{\mathcal{A}})$		$+2$	$+1$	0	-4	-3
$\mathcal{T}$	$\mathtt{WZ\overline{M}}$	$AA_{\mathcal{AB}} + BB_{\mathcal{A}} + AB_{\mathcal{AB}}$	$(AB_{\mathcal{AB}}, AA_{\mathcal{BA}}, BB_{\mathcal{A}})$		$+1.5$	$+1$	-0.5	-3	-2
	$\mathtt{WW\overline{M}}$	$2 \times AA_{\mathcal{AB}} + BB_{\mathcal{A}}$	$(AA_{\mathcal{BA}}, BB_{\mathcal{A}}, AA_{\mathcal{AB}}, BB_{\varepsilon} \prec \Gamma_B)$	$\cup$	$+2$	$+1$	0	-3	-2
	$\mathtt{WN\underline{M}}$	$AA_{\mathcal{AB}} + BB_{\mathcal{B}} + AB_{\mathcal{BA}}$	$(AB_{\mathcal{BA}}, AA_{\mathcal{AB}}, BB_{\mathcal{B}})$		$+1.5$	$+1$	-0.5	-3	-2
	$\mathtt{WW\underline{M}}$	$2 \times AA_{\mathcal{AB}} + BB_{\mathcal{B}}$	$(AA_{\mathcal{AB}}, BB_{\mathcal{A}}, AA_{\mathcal{AB}}, BB_{\varepsilon} \prec \Gamma_B)$	$\cup$	$+2$	$+1$	0	-3	-2
	$\mathtt{MN\overline{W}}$	$BB_{\mathcal{AB}} + AA_{\mathcal{A}} + AB_{\mathcal{BA}}$	$(AB_{\mathcal{BA}}, AA_{\mathcal{A}}, BB_{\mathcal{AB}})$		$+1.5$	$+1$	-0.5	-3	-2
	$\mathtt{MM\overline{W}}$	$2 \times BB_{\mathcal{AB}} + AA_{\mathcal{A}}$	$(BB_{\mathcal{BA}}, AA_{\mathcal{A}}, BB_{\mathcal{AB}}, AA_{\varepsilon} \prec \Gamma_A)$	$\cap$	$+2$	$+1$	0	-3	-2
	$\mathtt{MZ\underline{W}}$	$BB_{\mathcal{AB}} + AA_{\mathcal{B}} + AB_{\mathcal{AB}}$	$(AB_{\mathcal{AB}}, AA_{\mathcal{B}}, BB_{\mathcal{BA}})$		$+1.5$	$+1$	-0.5	-3	-2
	$\mathtt{MM\underline{W}}$	$2 \times BB_{\mathcal{AB}} + AA_{\mathcal{B}}$	$(BB_{\mathcal{AB}}, AA_{\mathcal{B}}, BB_{\mathcal{BA}}, AA_{\varepsilon} \prec \Gamma_A)$	$\cap$	$+2$	$+1$	0	-3	-2
$\mathcal{S}$	$\mathtt{ZN}$	$AB_{\mathcal{AB}} + AB_{\mathcal{BA}}$	$(AB_{\mathcal{AB}}, AB_{\mathcal{BA}})$		$+1$	$+1$	-1	-2	-1
	$\mathtt{\overline{WM}}$	$AA_{\mathcal{A}} + BB_{\mathcal{A}}$	$(AA_{\mathcal{A}}, BB_{\mathcal{A}})$		$+1$	$+1$	0	-1	-1
	$\mathtt{\underline{WM}}$	$AA_{\mathcal{B}} + BB_{\mathcal{B}}$	$(AA_{\mathcal{B}}, BB_{\mathcal{B}})$		$+1$	$+1$	0	-1	-1
	$\mathtt{W\overline{M}}$	$AA_{\mathcal{AB}} + BB_{\mathcal{A}}$	$(AA_{\mathcal{BA}}, BB_{\mathcal{A}})$		$+1$	$+1$	0	-1	-1
	$\mathtt{W\underline{M}}$	$AA_{\mathcal{AB}} + BB_{\mathcal{B}}$	$(AA_{\mathcal{AB}}, BB_{\mathcal{B}})$		$+1$	$+1$	0	-1	-1
	$\mathtt{WZ}$	$AA_{\mathcal{AB}} + AB_{\mathcal{AB}}$	$(AA_{\mathcal{BA}}, BB_{\varepsilon} \prec \Gamma_B, AB_{\mathcal{AB}})$	$\cup$	$+1.5$	$+1$	-0.5	-2	-1
	$\mathtt{WN}$	$AA_{\mathcal{AB}} + AB_{\mathcal{BA}}$	$(AA_{\mathcal{AB}}, BB_{\varepsilon} \prec \Gamma_B, AB_{\mathcal{BA}})$	$\cup$	$+1.5$	$+1$	-0.5	-2	-1
	$\mathtt{WW}$	$AA_{\mathcal{AB}} + AA_{\mathcal{AB}}$	$(AA_{\mathcal{AB}}, BB_{\varepsilon} \prec \Gamma_B, AA_{\mathcal{BA}}, BB_{\varepsilon} \prec \Gamma_B)$	$\cup$	$+2$	$+1$	0	-2	-1
	$\mathtt{M\overline{W}}$	$BB_{\mathcal{AB}} + AA_{\mathcal{A}}$	$(AA_{\mathcal{A}}, BB_{\mathcal{AB}})$		$+1$	$+1$	0	-1	-1
	$\mathtt{M\underline{W}}$	$BB_{\mathcal{AB}} + AA_{\mathcal{B}}$	$(AA_{\mathcal{B}}, BB_{\mathcal{BA}})$		$+1$	$+1$	0	-1	-1
	$\mathtt{MZ}$	$BB_{\mathcal{AB}} + AB_{\mathcal{AB}}$	$(BB_{\mathcal{BA}}, AB_{\mathcal{AB}}, AA_{\varepsilon} \prec \Gamma_A)$	$\cap$	$+1.5$	$+1$	-0.5	-2	-1
	$\mathtt{MN}$	$BB_{\mathcal{AB}} + AB_{\mathcal{BA}}$	$(BB_{\mathcal{AB}}, AB_{\mathcal{BA}}, AA_{\varepsilon} \prec \Gamma_A)$	$\cap$	$+1.5$	$+1$	-0.5	-2	-1
	$\mathtt{MM}$	$BB_{\mathcal{AB}} + BB_{\mathcal{AB}}$	$(BB_{\mathcal{AB}}, AA_{\varepsilon} \prec \Gamma_A, BB_{\mathcal{BA}}, AA_{\varepsilon} \prec \Gamma_A)$	$\cap$	$+2$	$+1$	0	-2	-1
$\mathcal{M}$	$\mathtt{ZZ\underline{W}\overline{M}}$	$2 \times AB_{\mathcal{AB}} + AA_{\mathcal{B}} + BB_{\mathcal{A}}$	$(AB_{\mathcal{AB}}, AA_{\mathcal{B}}, BA_{\mathcal{BA}}, BB_{\mathcal{A}})$		$+2$	$+1$	-1	-4	-2
	$\mathtt{NN\overline{W}\underline{M}}$	$2 \times AB_{\mathcal{BA}} + AA_{\mathcal{A}} + BB_{\mathcal{B}}$	$(AB_{\mathcal{BA}}, AA_{\mathcal{A}}, BA_{\mathcal{AB}}, BB_{\mathcal{B}})$		$+2$	$+1$	-1	-4	-2

Recap: Indels via Transitions



Recap: Indels via Transitions



$\lambda(C) = \frac{\aleph(C)}{2} + r(C)$

with $r(C) = \begin{cases} 1 & \text{if } C \text{ is indel-enclosing} \\ 0 & \text{otherwise} \end{cases}$

Transition counting in the ILP

Set label to 0 on active indel-edge in $\mathbb{A}$

$$r_v \leq 1 - x_{\{u,v\}} \qquad \forall \{u,v\} \in E_{ID}(\mathbb{A})\,,$$

Transition counting in the ILP

Set label to 0 on active indel-edge in $\mathbb{A}$

$$r_v \leq 1 - x_{\{u,v\}} \qquad \forall \ \{u, v\} \in E_{ID}(\mathbb{A}) \, ,$$

Set label to 1 on active indel-edge in $\mathbb{B}$

$$r_{v'} \geq x_{\{u',v'\}} \qquad \forall \ \{u', v'\} \in E_{ID}(\mathbb{B})$$

Transition counting in the ILP

Set label to 0 on active indel-edge in $\mathbb{A}$

$$r_v \leq 1 - x_{\{u,v\}} \qquad \forall \{u,v\} \in E_{ID}(\mathbb{A}),$$

Set label to 1 on active indel-edge in $\mathbb{B}$

$$r_{v'} \geq x_{\{u',v'\}} \qquad \forall \{u',v'\} \in E_{ID}(\mathbb{B})$$

Record the transition in variable

$$t_{\{u,v\}} \geq r_v - r_u \qquad \forall \{u,v\} \in E$$

Transition counting in the ILP

Set label to 0 on active indel-edge in $\mathbb{A}$

$$r_v \leq 1 - x_{\{u,v\}} \qquad \forall\ \{u,v\} \in E_{ID}(\mathbb{A})\,,$$

Set label to 1 on active indel-edge in $\mathbb{B}$

$$r_{v'} \geq x_{\{u',v'\}} \qquad \forall\ \{u',v'\} \in E_{ID}(\mathbb{B})$$

Record the transition in variable

$$t_{\{u,v\}} \geq r_v - r_u - (1 - x_{\{u,v\}}) \qquad \forall\ \{u,v\} \in E$$

What about $r(C)$?

$$
\begin{aligned}
w(Q[S,P]) = |\mathcal{C}^Q| - \sum_{C \in \mathcal{C}^Q \cup \mathcal{S}^Q} \left(\frac{\aleph(C)}{2} + r(C)\right) &= |\mathcal{C}^Q| - \frac{\aleph(Q)}{2} - \sum_{C \in \mathcal{C}^Q \cup \mathcal{S}^Q} r(C) \\
&= |\mathcal{C}^Q| - \frac{\aleph(Q)}{2} - |\{C \in \mathcal{C}^Q : C \text{ is indel-enclosing}\}| - |\mathcal{S}^Q| \\
&= |\{C \in \mathcal{C}^Q : C \text{ is not indel-enclosing}\}| - \frac{\aleph(Q)}{2} - |\mathcal{S}^Q|
\end{aligned}
$$

where $\mathcal{S}^Q$ are circular singletons in the decompostion,

$$
r(C) = \begin{cases} 1 & \text{if } C \text{ is indel-enclosing} \\ 0 & \text{otherwise} \end{cases}
$$

Removing indel-enclosing cycles from the count

Idea: Set the cycle label to 0.

$$\ell_i \leq i(1 - x_{\{v_i,v_j\}}) \quad \forall \, \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$$

Counting circular singletons

Idea: Each circular chromosome $k \in K$ is a potential circular singleton.

$$\sum_{e \in E_{ID}(k)} x_e - |k| + 1 \leq s_k \quad \forall k \in K$$

Final Objective Function

$$w(Q[S,P]) = |\{C \in \mathcal{C}^Q : C \text{ is not indel-enclosing}\}| - \frac{\aleph(Q)}{2} - |\mathcal{S}^Q|$$

Objective:

$$\texttt{Maximize} \sum_{1 \le i \le |V|} z_i - \frac{1}{2} \sum_{e \in E} t_e - \sum_{k \in K} s_k$$

Quiz

Match the following parts of the ILP to their function!

A	$\ell_i \leq i(1 - x_{\{v_i, v_j\}})$	$\forall \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
B	$r_v \leq 1 - x_{\{u,v\}}$ $r_{v'} \geq x_{\{u',v'\}}$	$\forall \{u, v\} \in E_{ID}(\mathbb{A})$, $\forall \{u', v'\} \in E_{ID}(\mathbb{B})$
C	$t_{\{u,v\}} \geq r_v - r_u - (1 - x_{\{u,v\}})$	$\forall \{u, v\} \in E$
D	$\sum_{e \in E_{id}^k} x_e - \lvert k \rvert + 1 \leq s_k$	$\forall k \in K$

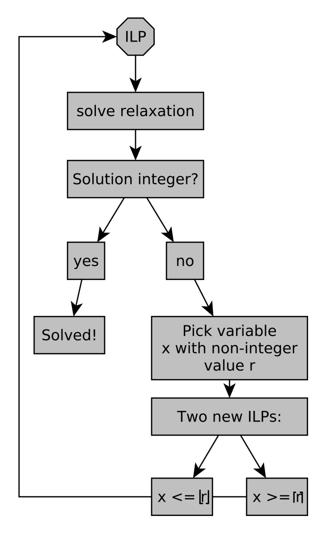
1 Setting run-variable (preparing to find transitions)

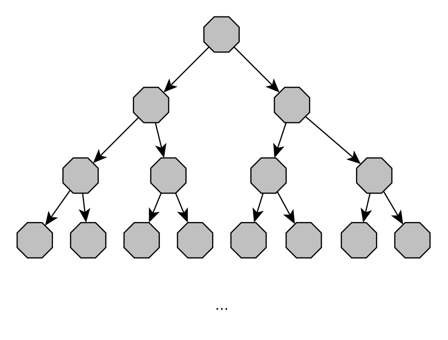
2 Removal of indel-enclosing cycles

3 Recording transitions

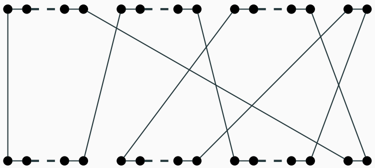
4 Flagging circular singletons

Excursus: ILP solvers





Refinement - Restricting where transitions occurr



Only permit transitions in adjacencies in $\mathbb{A}$

$$t_e = 0 \qquad \forall\ e \in E \setminus E_\alpha(\mathbb{A})$$

Only permit transitions next to indels

$$\sum_{\substack{d \in E_{ID}(\mathbb{A}),\\ d \cap e \neq \varnothing}} x_d - t_e \geq 0 \qquad \forall\ e \in E_\alpha(\mathbb{A})$$

Full ILP Solution

Objective:

$$\texttt{Maximize} \sum_{1 \leq i \leq |V|} z_i - \frac{1}{2} \sum_{e \in E} t_e - \sum_{k \in K} s_k$$

Constraints:

(C.01)	$x_e = 1$	$\forall\ e \in E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$
(C.02)	$\sum_{\{u,v\} \in E} x_{\{u,v\}} = 2$	$\forall\ u \in V$
(C.03)	$x_e = x_d$	$\forall\ e, d \in E_\xi$ such that e and d are siblings
(C.04)	$\ell_i \leq \ell_j + i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E$,
(C.06)	$i \cdot z_i \leq \ell_i$	$\forall\ 1 \leq i \leq \|V\|$
(C.05)	$\ell_i \leq i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
(C.07)	$r_v \leq 1 - x_{\{u,v\}}$	$\forall\ \{u, v\} \in E_{ID}(\mathbb{A})$,
	$r_{v'} \geq x_{\{u',v'\}}$	$\forall\ \{u', v'\} \in E_{ID}(\mathbb{B})$
(C.08)	$t_{\{u,v\}} \geq r_v - r_u - (1 - x_{\{u,v\}})$	$\forall\ \{u, v\} \in E$
(C.09)	$\sum_{d \in E_{ID}(\mathbb{A}),\ d \cap e \neq \varnothing} x_d - t_e \geq 0$	$\forall\ e \in E_\alpha(\mathbb{A})$
(C.10)	$t_e = 0$	$\forall\ e \in E \setminus E_\alpha(\mathbb{A})$
(C.11)	$\sum_{e \in E_{id}^k} x_e - \|k\| + 1 \leq s_k$	$\forall k \in K$

Domains:

(D.01)	$x_e \in \{0, 1\}$	$\forall\ e \in E$
(D.02)	$0 \leq \ell_i \leq i$	$\forall\ 1 \leq i \leq \|V\|$
(D.03)	$z_i \in \{0, 1\}$	$\forall\ 1 \leq i \leq \|V\|$
(D.04)	$r_v \in \{0, 1\}$	$\forall\ v \in V$
(D.05)	$t_e \in \{0, 1\}$	$\forall\ e \in E$
(D.06)	$s_k \in \{0, 1\}$	$\forall\ k \in K$

Full ILP Solution

Objective:

$$\texttt{Maximize} \sum_{1 \le i \le |V|} z_i - \frac{1}{2} \sum_{e \in E} t_e - \sum_{k \in K} s_k$$

Constraints:

(C.01)	$x_e = 1$	$\forall\ e \in E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$
(C.02)	$\sum_{\{u,v\} \in E} x_{\{u,v\}} = 2$	$\forall\ u \in V$
(C.03)	$x_e = x_d$	$\forall\ e, d \in E_\epsilon$ such that e and d are siblings
(C.04)	$\ell_i \le \ell_j + i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E$,
(C.06)	$i \cdot z_i \le \ell_i$	$\forall\ 1 \le i \le \|V\|$
(C.05)	$\ell_i \le i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
(C.07)	$r_v \le 1 - x_{\{u,v\}}$	$\forall\ \{u, v\} \in E_{ID}(\mathbb{A})$,
	$r_{v'} \ge x_{\{u',v'\}}$	$\forall\ \{u', v'\} \in E_{ID}(\mathbb{B})$
(C.08)	$t_{\{u,v\}} \ge r_v - r_u - (1 - x_{\{u,v\}})$	$\forall\ \{u, v\} \in E$
(C.09)	$\sum_{d \in E_{ID}(\mathbb{A}),\ d \cap e \ne \emptyset} x_d - t_e \ge 0$	$\forall\ e \in E_\alpha(\mathbb{A})$
(C.10)	$t_e = 0$	$\forall\ e \in E \setminus E_\alpha(\mathbb{A})$
(C.11)	$\sum_{e \in E_{id}^k} x_e - \|k\| + 1 \le s_k$	$\forall k \in K$

Domains:

(D.01)	$x_e \in \{0, 1\}$	$\forall\ e \in E$
(D.02)	$0 \le \ell_i \le i$	$\forall\ 1 \le i \le \|V\|$
(D.03)	$z_i \in \{0, 1\}$	$\forall\ 1 \le i \le \|V\|$
(D.04)	$r_v \in \{0, 1\}$	$\forall\ v \in V$
(D.05)	$t_e \in \{0, 1\}$	$\forall\ e \in E$
(D.06)	$s_k \in \{0, 1\}$	$\forall\ k \in K$

Shao et al.

Full ILP Solution

Objective:

$$\texttt{Maximize} \sum_{1 \le i \le |V|} z_i - \frac{1}{2} \sum_{e \in E} t_e - \sum_{k \in K} s_k$$

Constraints:

(C.01)	$x_e = 1$	$\forall\ e \in E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$
(C.02)	$\sum_{\{u,v\} \in E} x_{\{u,v\}} = 2$	$\forall\ u \in V$
(C.03)	$x_e = x_d$	$\forall\ e, d \in E_\xi$ such that e and d are siblings
(C.04)	$\ell_i \le \ell_j + i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E$,
(C.06)	$i \cdot z_i \le \ell_i$	$\forall\ 1 \le i \le \|V\|$
(C.05)	$\ell_i \le i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
(C.07)	$r_v \le 1 - x_{\{u,v\}}$	$\forall\ \{u, v\} \in E_{ID}(\mathbb{A})$,
	$r_{v'} \ge x_{\{u',v'\}}$	$\forall\ \{u', v'\} \in E_{ID}(\mathbb{B})$
(C.08)	$t_{\{u,v\}} \ge r_v - r_u - (1 - x_{\{u,v\}})$	$\forall\ \{u, v\} \in E$
(C.09)	$\sum_{d \in E_{ID}(\mathbb{A}),\ d \cap e \neq \emptyset} x_d - t_e \ge 0$	$\forall\ e \in E_\alpha(\mathbb{A})$
(C.10)	$t_e = 0$	$\forall\ e \in E \setminus E_\alpha(\mathbb{A})$
(C.11)	$\sum_{e \in E_{id}^k} x_e - \|k\| + 1 \le s_k$	$\forall k \in K$

Domains:

(D.01)	$x_e \in \{0, 1\}$	$\forall\ e \in E$
(D.02)	$0 \le \ell_i \le i$	$\forall\ 1 \le i \le \|V\|$
(D.03)	$z_i \in \{0, 1\}$	$\forall\ 1 \le i \le \|V\|$
(D.04)	$r_v \in \{0, 1\}$	$\forall\ v \in V$
(D.05)	$t_e \in \{0, 1\}$	$\forall\ e \in E$
(D.06)	$s_k \in \{0, 1\}$	$\forall\ k \in K$

DING extension

Full ILP Solution

Objective:

$$\texttt{Maximize} \sum_{1\leq i\leq |V|} z_i - \frac{1}{2}\sum_{e\in E} t_e - \sum_{k\in K} s_k$$

Constraints:

(C.01)	$x_e = 1$	$\forall\ e \in E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$
(C.02)	$\sum_{\{u,v\}\in E} x_{\{u,v\}} = 2$	$\forall\ u \in V$
(C.03)	$x_e = x_d$	$\forall\ e, d \in E_\xi$ such that e and d are siblings
(C.04)	$\ell_i \leq \ell_j + i(1 - x_{\{v_i,v_j\}})$	$\forall\ \{v_i, v_j\} \in E$,
(C.06)	$i \cdot z_i \leq \ell_i$	$\forall\ 1 \leq i \leq \lvert V\rvert$
(C.05)	$\ell_i \leq i(1 - x_{\{v_i,v_j\}})$	$\forall\ \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
(C.07)	$r_v \leq 1 - x_{\{u,v\}}$	$\forall\ \{u, v\} \in E_{ID}(\mathbb{A})$,
	$r_{v'} \geq x_{\{u',v'\}}$	$\forall\ \{u', v'\} \in E_{ID}(\mathbb{B})$
(C.08)	$t_{\{u,v\}} \geq r_v - r_u - (1 - x_{\{u,v\}})$	$\forall\ \{u, v\} \in E$
(C.09)	$\sum_{d\in E_{ID}(\mathbb{A}),\, d\cap e\neq\emptyset} x_d - t_e \geq 0$	$\forall\ e \in E_\alpha(\mathbb{A})$
(C.10)	$t_e = 0$	$\forall\ e \in E \setminus E_\alpha(\mathbb{A})$
(C.11)	$\sum_{e\in E^k_{id}} x_e - \lvert k\rvert + 1 \leq s_k$	$\forall k \in K$

Domains:

(D.01)	$x_e \in \{0, 1\}$	$\forall\ e \in E$
(D.02)	$0 \leq \ell_i \leq i$	$\forall\ 1 \leq i \leq \lvert V\rvert$
(D.03)	$z_i \in \{0, 1\}$	$\forall\ 1 \leq i \leq \lvert V\rvert$
(D.04)	$r_v \in \{0, 1\}$	$\forall\ v \in V$
(D.05)	$t_e \in \{0, 1\}$	$\forall\ e \in E$
(D.06)	$s_k \in \{0, 1\}$	$\forall\ k \in K$

Circ. singleton handling

Full ILP Solution

Objective:

$$\texttt{Maximize} \sum_{1 \le i \le |V|} z_i - \frac{1}{2} \sum_{e \in E} t_e - \sum_{k \in K} s_k$$

Constraints:

(C.01)	$x_e = 1$	$\forall\ e \in E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$
(C.02)	$\sum_{\{u,v\} \in E} x_{\{u,v\}} = 2$	$\forall\ u \in V$
(C.03)	$x_e = x_d$	$\forall\ e, d \in E_\xi$ such that e and d are siblings
(C.04)	$\ell_i \le \ell_j + i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E$,
(C.06)	$i \cdot z_i \le \ell_i$	$\forall\ 1 \le i \le \lvert V\rvert$
(C.05)	$\ell_i \le i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
(C.07)	$r_v \le 1 - x_{\{u,v\}}$	$\forall\ \{u, v\} \in E_{ID}(\mathbb{A})$,
	$r_{v'} \ge x_{\{u',v'\}}$	$\forall\ \{u', v'\} \in E_{ID}(\mathbb{B})$
(C.08)	$t_{\{u,v\}} \ge r_v - r_u - (1 - x_{\{u,v\}})$	$\forall\ \{u, v\} \in E$
(C.09)	$\sum_{d \in E_{ID}(\mathbb{A}),\ d \cap e \neq \emptyset} x_d - t_e \ge 0$	$\forall\ e \in E_\alpha(\mathbb{A})$
(C.10)	$t_e = 0$	$\forall\ e \in E \setminus E_\alpha(\mathbb{A})$
(C.11)	$\sum_{e \in E^k_{id}} x_e - \lvert k\rvert + 1 \le s_k$	$\forall k \in K$

Domains:

(D.01)	$x_e \in \{0, 1\}$	$\forall\ e \in E$
(D.02)	$0 \le \ell_i \le i$	$\forall\ 1 \le i \le \lvert V\rvert$
(D.03)	$z_i \in \{0, 1\}$	$\forall\ 1 \le i \le \lvert V\rvert$
(D.04)	$r_v \in \{0, 1\}$	$\forall\ v \in V$
(D.05)	$t_e \in \{0, 1\}$	$\forall\ e \in E$
(D.06)	$s_k \in \{0, 1\}$	$\forall\ k \in K$

Indel enclosing cycles

Full ILP Solution

Objective:

$$\texttt{Maximize} \sum_{1 \le i \le |V|} z_i - \frac{1}{2} \sum_{e \in E} t_e - \sum_{k \in K} s_k$$

Constraints:

(C.01)	$x_e = 1$	$\forall\ e \in E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$
(C.02)	$\sum_{\{u,v\} \in E} x_{\{u,v\}} = 2$	$\forall\ u \in V$
(C.03)	$x_e = x_d$	$\forall\ e, d \in E_\xi$ such that e and d are siblings
(C.04)	$\ell_i \le \ell_j + i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E$,
(C.06)	$i \cdot z_i \le \ell_i$	$\forall\ 1 \le i \le \|V\|$
(C.05)	$\ell_i \le i(1 - x_{\{v_i, v_j\}})$	$\forall\ \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
(C.07)	$r_v \le 1 - x_{\{u,v\}}$	$\forall\ \{u, v\} \in E_{ID}(\mathbb{A})$,
	$r_{v'} \ge x_{\{u',v'\}}$	$\forall\ \{u', v'\} \in E_{ID}(\mathbb{B})$
(C.08)	$t_{\{u,v\}} \ge r_v - r_u - (1 - x_{\{u,v\}})$	$\forall\ \{u, v\} \in E$
(C.09)	$\sum_{d \in E_{ID}(\mathbb{A}),\ d \cap e \neq \emptyset} x_d - t_e \ge 0$	$\forall\ e \in E_\alpha(\mathbb{A})$
(C.10)	$t_e = 0$	$\forall\ e \in E \setminus E_\alpha(\mathbb{A})$
(C.11)	$\sum_{e \in E^k_{id}} x_e - \|k\| + 1 \le s_k$	$\forall k \in K$

Domains:

(D.01)	$x_e \in \{0, 1\}$	$\forall\ e \in E$
(D.02)	$0 \le \ell_i \le i$	$\forall\ 1 \le i \le \|V\|$
(D.03)	$z_i \in \{0, 1\}$	$\forall\ 1 \le i \le \|V\|$
(D.04)	$r_v \in \{0, 1\}$	$\forall\ v \in V$
(D.05)	$t_e \in \{0, 1\}$	$\forall\ e \in E$
(D.06)	$s_k \in \{0, 1\}$	$\forall\ k \in K$

Transition counting

Full ILP Solution

Objective:

$$\texttt{Maximize} \sum_{1\leq i\leq |V|} z_i - \frac{1}{2}\sum_{e\in E} t_e - \sum_{k\in K} s_k$$

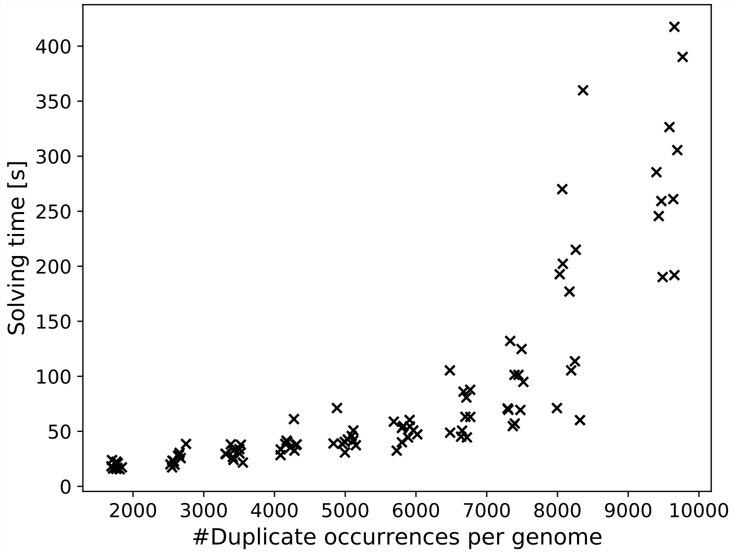
Constraints:

(C.01)	$x_e = 1$	$\forall\ e \in E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$
(C.02)	$\sum_{\{u,v\}\in E} x_{\{u,v\}} = 2$	$\forall\ u \in V$
(C.03)	$x_e = x_d$	$\forall\ e, d \in E_\xi$ such that e and d are siblings
(C.04)	$\ell_i \leq \ell_j + i(1 - x_{\{v_i,v_j\}})$	$\forall\ \{v_i, v_j\} \in E$,
(C.06)	$i \cdot z_i \leq \ell_i$	$\forall\ 1 \leq i \leq \lvert V\rvert$
(C.05)	$\ell_i \leq i(1 - x_{\{v_i,v_j\}})$	$\forall\ \{v_i, v_j\} \in E_{ID}(\mathbb{A}) \cup E_{ID}(\mathbb{B})$
(C.07)	$r_v \leq 1 - x_{\{u,v\}}$	$\forall\ \{u, v\} \in E_{ID}(\mathbb{A})$,
	$r_{v'} \geq x_{\{u',v'\}}$	$\forall\ \{u', v'\} \in E_{ID}(\mathbb{B})$
(C.08)	$t_{\{u,v\}} \geq r_v - r_u - (1 - x_{\{u,v\}})$	$\forall\ \{u, v\} \in E$
(C.09)	$\sum_{\substack{d\in E_{ID}(\mathbb{A}),\\ d\cap e\neq\emptyset}} x_d - t_e \geq 0$	$\forall\ e \in E_\alpha(\mathbb{A})$
(C.10)	$t_e = 0$	$\forall\ e \in E \setminus E_\alpha(\mathbb{A})$
(C.11)	$\sum_{e\in E^k_{id}} x_e - \lvert k\rvert + 1 \leq s_k$	$\forall k \in K$

Domains:

(D.01)	$x_e \in \{0, 1\}$	$\forall\ e \in E$
(D.02)	$0 \leq \ell_i \leq i$	$\forall\ 1 \leq i \leq \lvert V\rvert$
(D.03)	$z_i \in \{0, 1\}$	$\forall\ 1 \leq i \leq \lvert V\rvert$
(D.04)	$r_v \in \{0, 1\}$	$\forall\ v \in V$
(D.05)	$t_e \in \{0, 1\}$	$\forall\ e \in E$
(D.06)	$s_k \in \{0, 1\}$	$\forall\ k \in K$

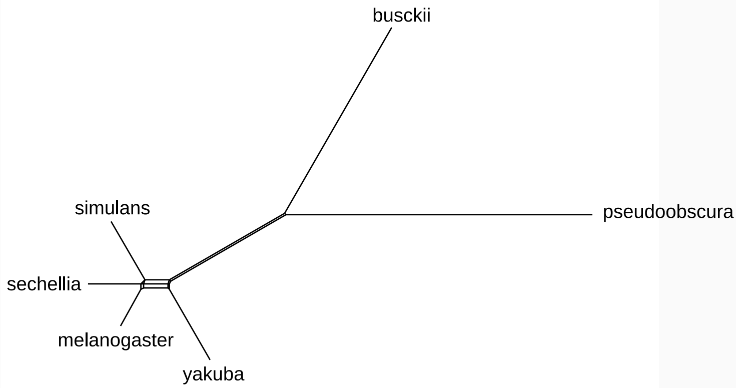
ILPs can be very fast



ILPs can be very fast

Genome pair	Max. multiplicity of dupl. marker	#duplicate markers	#duplicate occ.	d_{DCJ}^{id}	solving time [s]
dbus-dmel	23	303	832	4661	6.02
dbus-dpse	17	361	934	4688	5.29
dbus-dsec	15	295	766	4710	5.64
dbus-dsim	13	281	721	4767	5.05
dbus-dyak	19	318	785	4756	5.00
dmel-dpse	23	469	1319	3799	32218.93
dmel-dsec	23	326	902	901	6.78
dmel-dsim	23	322	893	1093	5.73
dmel-dyak	23	362	972	1379	7.22
dpse-dsec	17	464	1227	3866	13.82
dpse-dsim	17	449	1198	3962	6.81
dpse-dyak	19	481	1259	3951	8.96
dsec-dsim	15	314	843	1138	5.67
dsec-dyak	19	354	903	1516	6.56
dsim-dyak	19	347	864	1661	23.07

Resolved Phylogeny



Not quite, but for an improved procedure, stay tuned for next lecture :)

Literature

Gurobi mip solver introduction.
`https://www.gurobi.com/resource/mip-basics/`.
Accessed: 2021-01-26.

Bohnenkämper, L., Braga, M. D., Doerr, D., and Stoye, J. (0).
Computing the rearrangement distance of natural genomes.
Journal of Computational Biology, 0(0):null.
PMID: 33393848.