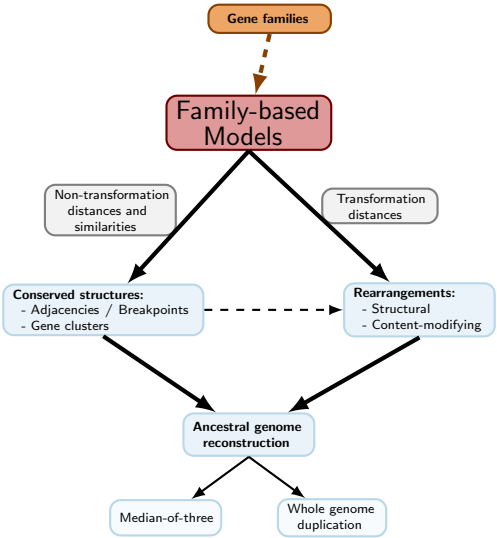


Topics of today:

1. Family-based \times Family-free setting
2. Family-free DCJ distance
3. Family-free DCJ-indel distance

Family-based setting



Are family assignments accurate?

Ideal situation:



Are family assignments accurate?

Ideal situation:

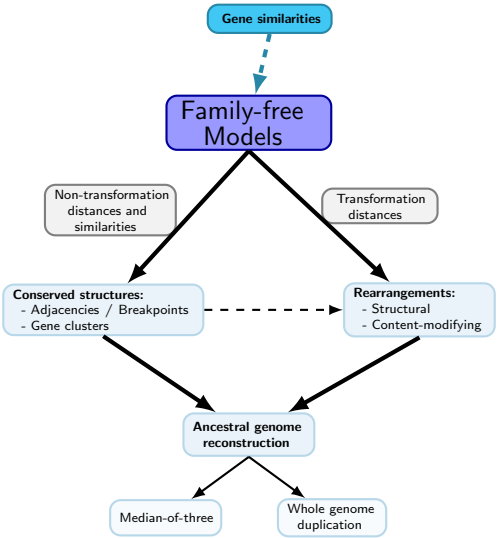


- ▶ Family assignments are most of the time made automatically
- ▶ Even in the absence of errors, there may be ambiguities:

Reality:

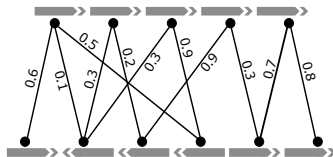


Alternative: family-free setting



Family-free DCJ distance

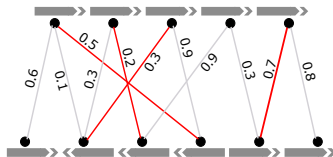
No family assignments , but pairwise normalized similarities
(above some threshold $x \in [0, 1]$)



$$x = 0.1$$

Family-free DCJ distance

No family assignments, but pairwise normalized similarities
(above some threshold $x \in [0, 1]$)



matching M

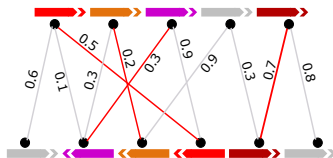
$$x = 0.1$$

$$|M| = 4$$

$$w(M) = 1.7$$

Family-free DCJ distance

No family assignments, but pairwise normalized similarities (above some threshold $x \in [0, 1]$)



$$x = 0.1$$

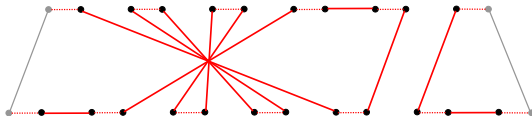
$$|M| = 4$$

$$w(M) = 1.7$$

matching $M \rightarrow$ singular mapped genomes \mathbb{A}^M and \mathbb{B}^M



capped relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



capping of canonical genomes
(ignores indels/recombinations)

Taking the weights into consideration

Weighted DCJ distance of mapped genomes

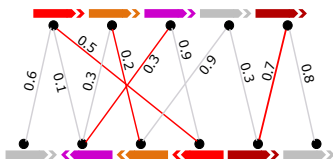
$$\begin{aligned} \text{wd}_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) &= d_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) + |M| - w(M) \\ &= p_* + |M| - |\mathcal{C}| + |M| - w(M) \end{aligned}$$

$|M| - w(M)$: penalizes edges of M with similarity < 1

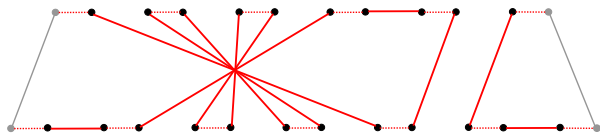
genes that are not covered by M are simply ignored

Weighted and unweighted DCJ distances of mapped genomes

$|M| = 4$ is maximal, $w(M) = 1.7$



M	$ M $	d_{DCJ}	$ M - w(M)$	wd_{DCJ}
M	4	1	2.3	3.3

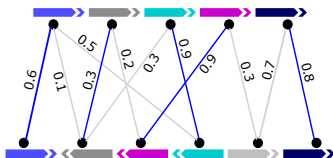


$$d_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 4 - 4 = 1$$

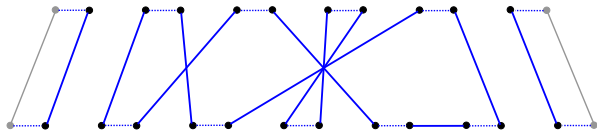
$$\text{wd}_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 2.3 = 3.3$$

Weighted and unweighted DCJ distances of mapped genomes

$|M| = 5$ is maximal, $w(M) = 3.5$



M	$ M $	d_{DCJ}	$ M - w(M)$	wd_{DCJ}
M	4	1	2.3	3.3
M	5	2	1.5	3.5

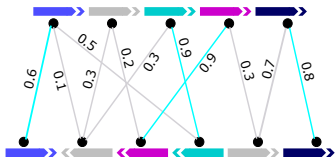


$$d_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 5 - 4 = 2$$

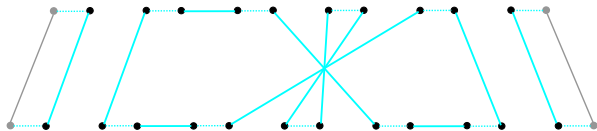
$$\text{wd}_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 2 + 1.5 = 3.5$$

Weighted and unweighted DCJ distances of mapped genomes

$|M| = 4$ is non-maximal, $w(M) = 3.2$



M	$ M $	d_{DCJ}	$ M - w(M)$	wd_{DCJ}
M	4	1	2.3	3.3
M	5	2	1.5	3.5
M	4	1	0.8	1.8

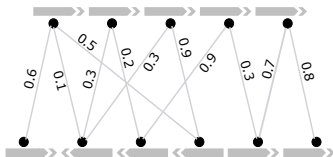


$$d_{\text{DCJ}}(A^M, B^M) = 1 + 4 - 4 = 1$$

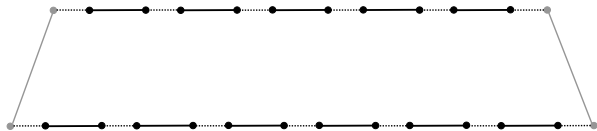
$$\text{wd}_{\text{DCJ}}(A^M, B^M) = 1 + 0.8 = 1.8$$

Weighted and unweighted DCJ distances of mapped genomes

M is empty, $w(M) = 0$



M	$ M $	d_{DCJ}	$ M - w(M)$	wd_{DCJ}
M	4	1	2.3	3.3
M	5	2	1.5	3.5
M	4	1	0.8	1.8
M	0	0	0	0

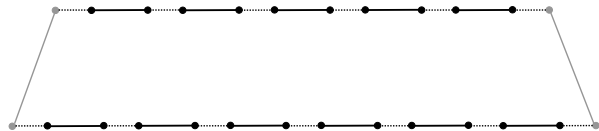
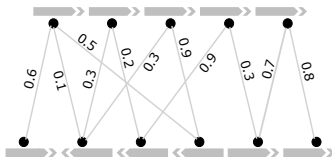


$$d_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 0 - 1 = 0$$

$$\text{wd}_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 0 + 0 = 0$$

Weighted and unweighted DCJ distances of mapped genomes

M is empty, $w(M) = 0$



$$d_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 0 - 1 = 0$$

$$wd_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 0 + 0 = 0$$

M	$ M $	d_{DCJ}	$ M - w(M)$	wd_{DCJ}
M	4	1	2.3	3.3
M	5	2	1.5	3.5
M	4	0	0.3	4.3
M	0	0	0	0
\vdots	\vdots	\vdots	\vdots	\vdots

Family-free DCJ distance

$$\min_{M \in \mathfrak{M}_{\text{MAX}}} \{ wd_{\text{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) \}$$

$\mathfrak{M}_{\text{MAX}}$: set of all maximal matchings

NP-hard

Quiz 1

Recap Shao-Lin-Moret: Match the parts of the ILP to their function!

A $l_i \leq l_j + i(1 - x_{\{v_i, v_j\}}) \quad \forall \{v_i, v_j\} \in E$

B $\sum_{\{u,v\} \in E} x_{\{u,v\}} = 2$

C $i \cdot c_i \leq l_i$

D $x_e = 1$

E $x_e = x_d$

$\forall u \in V$

$\forall 1 \leq i \leq |V|$

$\forall e \in E_\alpha^A \cup E_\alpha^B$

$\forall e, d \in E_\xi$ 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 smallest label

4 A decomposition consists only of simple cycles

5 Cycle labels of adjacent vertices are the same

ILP formulation for the family-free DCJ distance

We have a capped multi-relational graph, but here each gene can be potentially ignored:
⇒ each gene has an indel edge and ignoring a gene is done by selecting its indel edge

siblings have the same weight as the edge in the similarity graph

Besides variables and constraints from Shao-Lin-Moret, an extra constraint assures a maximal matching:

- at least one of the vertices connected by an edge in the similarity graph must be chosen
- at most one of the corresponding indel edges in the capped MRG can be selected:

$$\underline{x_a} + \underline{x_b} \leq \textcircled{1} \quad \left\{ \begin{array}{l} \forall a \in E_{ID}^A, b \in E_{ID}^B, \\ \text{siblings } d, e \in E_{\xi} \text{ and} \\ a \cap (d \cup e) \neq \emptyset \text{ and } b \cap (d \cup e) \neq \emptyset \\ e \cup b = d \cup e \end{array} \right.$$

Weighted DCJ distance formula

$$wd_{DCJ}(A^M, B^M) = p_* + |M| - |C| + |M| - w(M) = p_* + 2|M| - |C| - w(M)$$

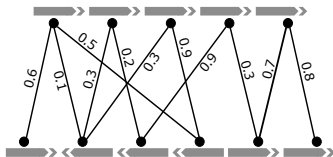
Objective function:

$$\min p_* + \sum_{e \in E_{\xi}} x_e - \sum_{1 \leq i \leq |V|} c_i - \frac{1}{2} \sum_{e \in E_{\xi}} w_e x_e$$

(one edge of the matching corresponds to a pair of edges in E_{ξ})

Family-free DCJ-indel distance

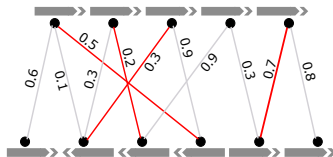
No family assignments , but pairwise normalized similarities
(above some threshold $x \in [0, 1]$)



$$x = 0.1$$

Family-free DCJ-indel distance

No family assignments, but pairwise normalized similarities (above some threshold $x \in [0, 1]$)



matching M

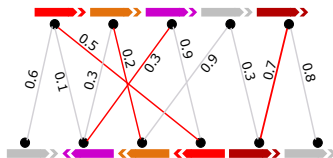
$$x = 0.1$$

$$|M| = 4$$

$$w(M) = 1.7$$

Family-free DCJ-indel distance

No family assignments, but pairwise normalized similarities
(above some threshold $x \in [0, 1]$)



$$x = 0.1$$

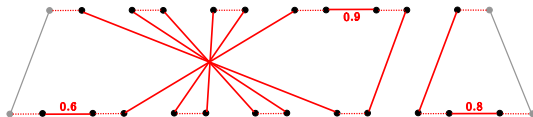
$$|M| = 4$$

$$w(M) = 1.7$$

matching $M \rightarrow$ singular mapped genomes A^M and B^M



weighted capped relational diagram



weight of indel edge = maximum similarity to the corresponding marker

capping of singular genomes
(with indels/recombinations)

\tilde{M} : set of indel edges
(complement of M)

$$w(\tilde{M}) = 2.3$$

Taking the weights into consideration

Weighted DCJ-indel distance of mapped genomes

$$\begin{aligned}\text{wd}_{\text{DCJ}}^{\text{ID}}(\mathbb{A}^M, \mathbb{B}^M) &= d_{\text{DCJ}}^{\text{ID}}(\mathbb{A}^M, \mathbb{B}^M) + \boxed{|M| - w(M)} + \boxed{w(\tilde{M})} \\ &= p_* + |M| - |\mathcal{C}| + \sum_{c \in \mathcal{C}_{\text{US}}} \lambda(c) + |M| - w(M) + w(\tilde{M}) \\ &= p_* + |M| - |\mathcal{C}| + |\mathcal{C}^r| + |\mathcal{S}| + \frac{\aleph}{2} + |M| - w(M) + w(\tilde{M}) \\ &= p_* + |M| - |\mathcal{C}^{\bar{r}}| + |\mathcal{S}| + \frac{\aleph}{2} + |M| - w(M) + w(\tilde{M})\end{aligned}$$

$|M| - w(M)$: penalizes edges of M with similarity < 1

$w(\tilde{M})$: penalizes markers of \tilde{M} with some similarity > 0

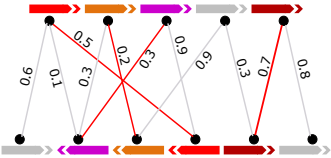
\mathcal{C}^r : set of indel-enclosing cycles

$\mathcal{C}^{\bar{r}}$: set of indel-free cycles

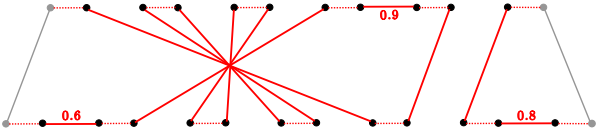
\mathcal{S} : set of circular singletons

Weighted and unweighted DCJ-indel distances of mapped genomes

$|M| = 4$ is maximal, $w(M) = 1.7$, $w(\tilde{M}) = 2.3$



M	$ M $	d_{DCJ}^{ID}	$ M - w(M)$	$w(\tilde{M})$	wd_{DCJ}^{ID}
M	4	4	2.3	2.3	8.6

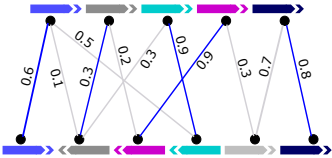


$$d_{DCJ}^{ID}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 4 - 2 + 0 + \frac{2}{2} = 4$$

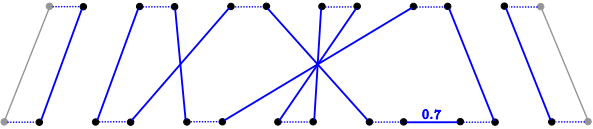
$$wd_{DCJ}^{ID}(\mathbb{A}^M, \mathbb{B}^M) = 4 + 4 - 1.7 + 2.3 = 8.6$$

Weighted and unweighted DCJ-indel distances of mapped genomes

$|M| = 5$ is maximal, $w(M) = 3.5$, $w(\tilde{M}) = 0.7$



M	$ M $	d_{DCJ}^{ID}	$ M - w(M)$	$w(\tilde{M})$	wd_{DCJ}^{ID}
M	4	4	2.3	2.3	8.6
M	5	3	1.5	0.7	5.2

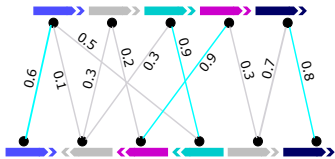


$$d_{DCJ}^{ID}(A^M, B^M) = 1 + 5 - 3 + 0 + 0 = 3$$

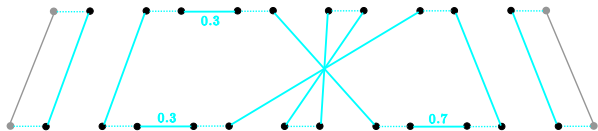
$$wd_{DCJ}^{ID}(A^M, B^M) = 3 + 5 - 3.5 + 0.7 = 5.2$$

Weighted and unweighted DCJ-indel distances of mapped genomes

$|M| = 4$ is non-maximal, $w(M) = 3.2$, $w(\tilde{M}) = 1.3$



M	$ M $	$d_{\text{DCJ}}^{\text{ID}}$	$ M - w(M)$	$w(\tilde{M})$	$\text{wd}_{\text{DCJ}}^{\text{ID}}$
M	4	4	2.3	2.3	8.6
M	5	3	1.5	0.7	5.2
M	4	3	0.8	1.3	5.1

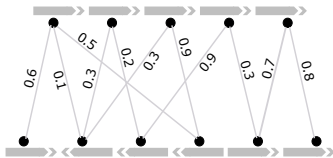


$$d_{\text{DCJ}}^{\text{ID}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 4 - 3 + 0 + \frac{2}{2} = 3$$

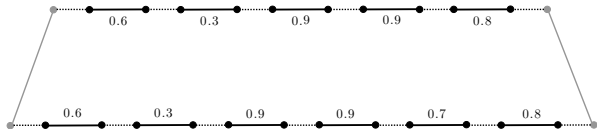
$$\text{wd}_{\text{DCJ}}^{\text{ID}}(\mathbb{A}^M, \mathbb{B}^M) = 3 + 4 - 3.2 + 1.3 = 5.1$$

Weighted and unweighted DCJ-indel distances of mapped genomes

M is empty, $w(M) = 0$, $w(\tilde{M}) = 7.7$



M	$ M $	$d_{\text{DCJ}}^{\text{ID}}$	$ M - w(M)$	$w(\tilde{M})$	$\text{wd}_{\text{DCJ}}^{\text{ID}}$
\tilde{M}	4	4	2.3	2.3	8.6
M	5	3	1.5	0.7	5.2
\tilde{M}	4	3	0.8	1.3	5.1
M	0	2	0	7.7	9.7

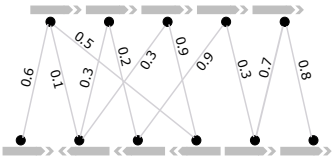


$$d_{\text{DCJ}}^{\text{ID}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 0 - 0 + 0 + \frac{2}{2} = 2$$

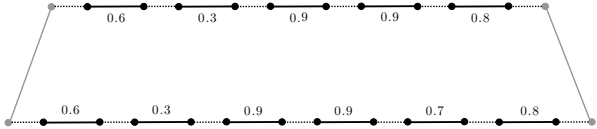
$$\text{wd}_{\text{DCJ}}^{\text{ID}}(\mathbb{A}^M, \mathbb{B}^M) = 2 + 0 - 0 + 7.7 = 9.7$$

Weighted and unweighted DCJ-indel distances of mapped genomes

M is empty, $w(M) = 0$, $w(\tilde{M}) = 7.7$



M	$ M $	d_{DCJ}^{ID}	$ M - w(M)$	$w(\tilde{M})$	wd_{DCJ}^{ID}
\tilde{M}	4	4	2.3	2.3	8.6
M	5	3	1.5	0.7	5.2
\tilde{M}	4	3	0.8	1.3	5.1
M	0	2	0	7.7	9.7
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots



$$d_{DCJ}^{ID}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 0 - 0 + 0 + \frac{2}{2} = 2$$

$$wd_{DCJ}^{ID}(\mathbb{A}^M, \mathbb{B}^M) = 2 + 0 - 0 + 7.7 = 9.7$$

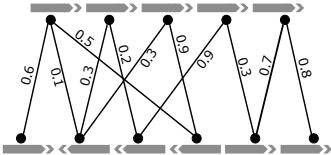
Family-free DCJ-indel distance

$$\min_{M \in \mathfrak{M}} \{ wd_{DCJ}^{ID}(\mathbb{A}^M, \mathbb{B}^M) \}$$

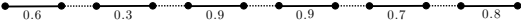
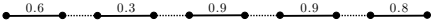
\mathfrak{M} : set of all matchings

NP-hard

Approach for ILP

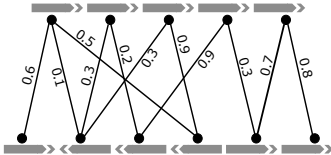


Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$

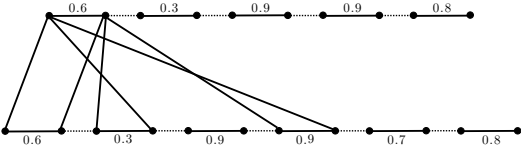


each marker has its **weighted indel edge** (weight = max. similarity)

Approach for ILP



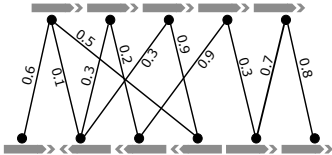
Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



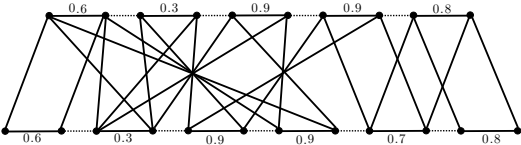
sibling weights are omitted

each marker has its **weighted indel edge** (weight = max. similarity)

Approach for ILP



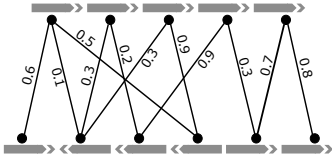
Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



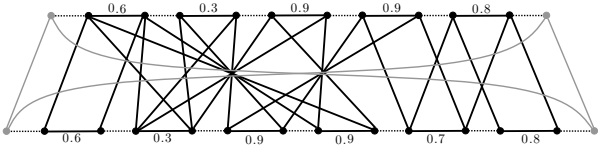
sibling weights are omitted

each marker has its **weighted indel edge** (weight = max. similarity)

Approach for ILP



Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



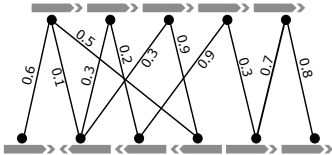
sibling weights are omitted

each marker has its **weighted indel edge** (weight = max. similarity)

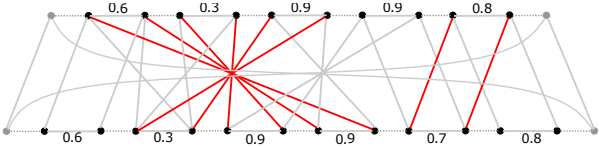
add $2p_*$ vertices (cap extremities)

link each cap extremity in genome \mathbb{A}
to each cap extremity in genome \mathbb{B}

Approach for ILP



Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$

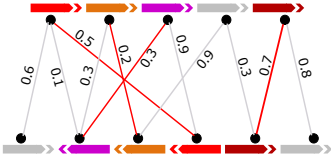


sibling weights are omitted

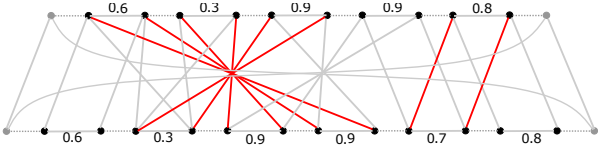
matching of extremity edges:

sibling set S
(pairs of siblings)

Approach for ILP



Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



sibling weights are omitted

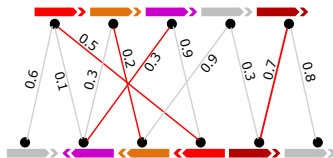
$$|S| = 2|M|$$

$$w(S) = 2w(M)$$

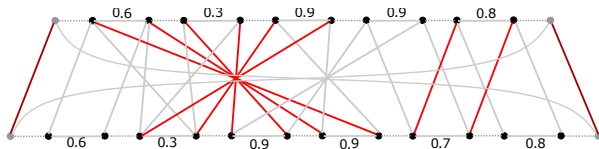
matching of extremity edges:

sibling set $S \rightarrow M$
(pairs of siblings)

Approach for ILP



Capped family-free relational diagram : $\rho_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



sibling weights are omitted

$$|S| = 2|M|$$

$$w(S) = 2w(M)$$

matching of extremity edges:

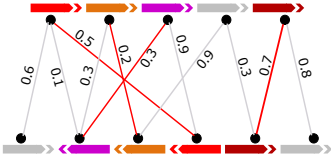
sibling set $S \rightarrow M$

(pairs of siblings)

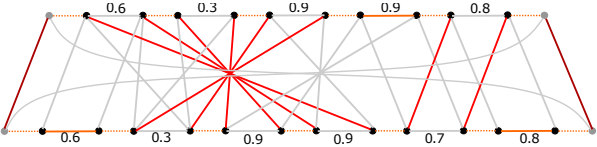
maximal capping set P

(covers all cap extremities)

Approach for ILP



Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



sibling weights are omitted

$$|S| = 2|M|$$

$$w(S) = 2w(M)$$

matching of extremity edges:

sibling set $S \rightarrow M$
(pairs of siblings)

maximal capping set P
(covers all cap extremities)

capped consistent decomposition
 $Q[S, P]$

$$\left\{ \begin{array}{l} S \cup P \\ \text{all adjacency edges} \\ \text{complement } \tilde{S} \equiv \tilde{M} \end{array} \right.$$

Optimization formula

DCJ-indel distance of a capped consistent decomposition

$$d_{\text{DCJ}}^{\text{ID}}(Q[S, P]) = p_* + \boxed{\frac{|S|}{2}} - |\mathcal{C}^{\tilde{r}}| + |S| + \frac{\aleph}{2}$$

$$\boxed{|S| = 2|M|}$$

Optimization formula

DCJ-indel distance of a capped consistent decomposition

$$d_{\text{DCJ}}^{\text{ID}}(Q[S, P]) = p_* + \boxed{\frac{|S|}{2}} - |\mathcal{C}^{\tilde{r}}| + |S| + \frac{\aleph}{2}$$

$$\boxed{|S| = 2|M|}$$

Weighted DCJ-indel distance of a capped consistent decomposition

$$\begin{aligned} wd_{\text{DCJ}}^{\text{ID}}(Q[S, P]) &= d_{\text{DCJ}}^{\text{ID}}(Q[S, P]) + \boxed{\frac{|S|}{2} - \frac{w(S)}{2}} + w(\tilde{S}) \\ &= p_* + \frac{|S|}{2} - |\mathcal{C}^{\tilde{r}}| + |S| + \frac{\aleph}{2} + \frac{|S|}{2} - \frac{w(S)}{2} + w(\tilde{S}) \\ &= p_* + |S| - |\mathcal{C}^{\tilde{r}}| + |S| + \frac{\aleph}{2} - \frac{w(S)}{2} + w(\tilde{S}) \end{aligned}$$

$$\boxed{w(S) = 2w(M)}$$

Optimization formula

DCJ-indel distance of a capped consistent decomposition

$$d_{\text{DCJ}}^{\text{ID}}(Q[S, P]) = p_* + \boxed{\frac{|S|}{2}} - |\mathcal{C}^{\tilde{r}}| + |S| + \frac{\aleph}{2}$$

$$\boxed{|S| = 2|M|}$$

Weighted DCJ-indel distance of a capped consistent decomposition

$$\begin{aligned} wd_{\text{DCJ}}^{\text{ID}}(Q[S, P]) &= d_{\text{DCJ}}^{\text{ID}}(Q[S, P]) + \boxed{\frac{|S|}{2} - \frac{w(S)}{2}} + w(\tilde{S}) \\ &= p_* + \frac{|S|}{2} - |\mathcal{C}^{\tilde{r}}| + |S| + \frac{\aleph}{2} + \frac{|S|}{2} - \frac{w(S)}{2} + w(\tilde{S}) \\ &= p_* + |S| - |\mathcal{C}^{\tilde{r}}| + |S| + \frac{\aleph}{2} - \frac{w(S)}{2} + w(\tilde{S}) \end{aligned}$$

$$\boxed{w(S) = 2w(M)}$$

Family-free DCJ-indel distance

$$\min_{S \in \mathfrak{S}, P \in \mathfrak{P}_{\text{MAX}}} \{ wd_{\text{DCJ}}^{\text{ID}}(Q[S, P]) \}$$


\mathfrak{S} : set of sibling sets

$\mathfrak{P}_{\text{MAX}}$: set of maximal capping sets

Quiz 2

Recap natural genomes: Match the following parts of the ILP to their function!

A $l_i \leq i(1 - x_{\{v_i, v_j\}})$



B $r_v \leq 1 - x_{\{u, v\}}$
 $r_{v'} \geq x_{\{u', v'\}}$

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

D $\sum_{e \in E_{id}^k} x_e - |k| + 1 \leq s_k$

$\forall \{v_i, v_j\} \in E_{ID}^A \cup E_{ID}^B$

$\forall \{u, v\} \in E_{ID}^A,$
 $\forall \{u', v'\} \in E_{ID}^B$

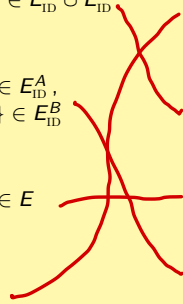
$\forall k \in K$

1 Flagging circular singletons

2 Removal of indel-enclosing cycles

3 Recording transitions

4 Setting run-variable
(preparing to find transitions)



ILP formulation for the family-free DCJ-indel distance

Previous formulations:

DCJ distance of balanced genomes (Shao *et al.*, 2014)

DCJ-indel distance of natural genomes (Bohnenkämper *et al.*, 2020)

Selecting a consistent decomposition:

(Shao *et al.*, 2014)

$$x_a = 1 \quad \forall a \in E_\alpha^A \cup E_\alpha^B$$

$$\sum_{uv \in E} x_{uv} = 2 \quad \forall u \in V$$

$$x_e = x_d \quad \forall e, d \in E_\xi, e, d \text{ are siblings}$$

ILP formulation for the family-free DCJ-indel distance

Previous formulations:

DCJ distance of balanced genomes (Shao *et al.*, 2014)

DCJ-indel distance of natural genomes (Bohnenkämper *et al.*, 2020)

Counting indel-free cycles:

(Shao *et al.*, 2014)

(adapted by Bohnenkämper *et al.*, 2020)

$$\left. \begin{array}{l} \ell_i \leq \ell_j + i(1 - x_{v_i v_j}) \\ \ell_j \leq \ell_i + j(1 - x_{v_i v_j}) \end{array} \right\} \forall v_i v_j \in E$$

$$\left. \begin{array}{l} \ell_i \leq i(1 - x_{v_i v_j}) \\ \ell_j \leq j(1 - x_{v_i v_j}) \end{array} \right\} \forall v_i v_j \in E_{ID}^A \cup E_{ID}^B$$

$$i \cdot c_i \leq \ell_i \quad \forall 1 \leq i \leq |V|$$

ILP formulation for the family-free DCJ-indel distance

Previous formulations:

DCJ distance of balanced genomes (Shao *et al.*, 2014)

DCJ-indel distance of natural genomes (Bohnenkämper *et al.*, 2020)

Counting singletons:

(Bohnenkämper *et al.*, 2020)

$$\sum_{e \in E_{\text{ID}}^k} x_e - |k| \leq s_k \quad \forall k \in K$$

ILP formulation for the family-free DCJ-indel distance

Previous formulations:

DCJ distance of balanced genomes (Shao *et al.*, 2014)

DCJ-indel distance of natural genomes (Bohnenkämper *et al.*, 2020)

Counting transitions:

(Bohnenkämper *et al.*, 2020)

$$\left. \begin{array}{l} r_v \leq 1 - x_{uv} \\ r_{v'} \geq x_{u'v'} \end{array} \right\} \quad \begin{array}{l} \forall uv \in E_{\text{ID}}^{\text{A}} \\ \forall u'v' \in E_{\text{ID}}^{\text{B}} \end{array}$$

$$\left. \begin{array}{l} t_{uv} \geq r_v - r_u - (1 - x_{uv}) \\ t_{uv} \geq r_u - r_v - (1 - x_{uv}) \end{array} \right\} \quad \forall uv \in E$$

$$\sum_{\substack{d \in E_{\text{ID}}^{\text{A}} \\ d \cap e \neq \emptyset}} x_d - t_a \geq 0 \quad \forall a \in E_{\alpha}^{\text{A}}$$

$$t_e = 0 \quad \forall e \in E \setminus E_{\alpha}^{\text{A}}$$

ILP formulation for the family-free DCJ-indel distance

Weighted DCJ-indel distance formula

$$wd_{\text{DCJ}}^{\text{ID}}(Q[S, P]) = p_* + |S| - |C^{\tilde{r}}| + |S| + \frac{N}{2} - \frac{w(S)}{2} + w(\tilde{S})$$

Objective function:

$$\min p_* + \sum_{e \in E_{\xi}} x_e + \sum_{1 \leq i \leq |V|} c_i + \sum_{k \in K} s_k + \frac{1}{2} \sum_{a \in E} t_a - \frac{1}{2} \sum_{e \in E_{\xi}} w_e x_e + \sum_{c \in E_{\text{ID}}} w_c x_c$$

Available at

<https://gitlab.ub.uni-bielefeld.de/gi/gen-diff>

Running times (or gap in %) for CPLEX with max. CPU time of 3h

Pairwise comparisons of *Drosophila* genomes

~ 13,000 genes per genome, distributed in 5-6 chromosomes

gene similarities obtained using FFGC (Doerr *et al.*, 2018):

considering all similarities that are strictly greater than $x = 0$, the pairwise similarity graphs have an average of 11.2 and at most 95 connections per gene.

DIFF on similarity graphs of $x = 0.3$, with an average of 1.92 and at most 31 connections per gene:

species	<i>pseudoobscura</i>	<i>sechellia</i>	<i>simulans</i>	<i>yakuba</i>	<i>busckii</i>
<i>melanogaster</i>	0.76%	4,431.78s	109.60s	201.49s	540.19s
<i>pseudoobscura</i>		163.12s	764.24s	5,782.73s	290.12s
<i>sechellia</i>			103.33s	146.88s	415.23s
<i>simulans</i>				216.77s	115.54s
<i>yakuba</i>					153.36s

(3h=10,800s):

DING on OMA families with an average of 1.04 and at most 23 occurrences:

All comparisons finished very fast, ranging from 2 to 32 seconds.

Comparing DIFF and DING on CPLEX with max. CPU time of 3h

Balancing the number of multiple connections in both models:

Extending the connected components of similarity graphs to cliques



DING on families derived from similarity graphs extended cliques:

All but one comparisons reached the time limit of 3h.

DIFF on similarity graphs with extended cliques (new edges received weight=0.3):

Only one comparison reached the time limit of 3h, the others took 380 seconds on average.

Observation: D_{ING} $\left\{ \begin{array}{l} \text{has a smaller search space only composed of maximal sibling-sets} \\ \text{running times were considerably longer} \end{array} \right.$

Probable explanation:

There is a larger number of co-optimal solutions in the DCJ-indel distance of natural genomes.

In the family-free DCJ-indel model the co-optimality is reduced by weights, allowing **DIFF** to converge faster.

⇒ Indeed, in a simulation in which the weights of all edges of the similarity graphs were set to 1, the running times of **DIFF** were much slower than those of **DING** for instances with the same number of multiple connections.

Gene homologies established by DIFF compared to Flybase

Flybase (flybase.org) established gene families (homolog gene sets) for the three species

$\left\{ \begin{array}{l} D. melanogaster \\ D. simulans \\ D. yakuba \end{array} \right.$

Classification of pairs of homologous genes inferred for these three species with DIFF (for $x = 0.3$):

Match: (97.3%) both genes are in the same Flybase family;

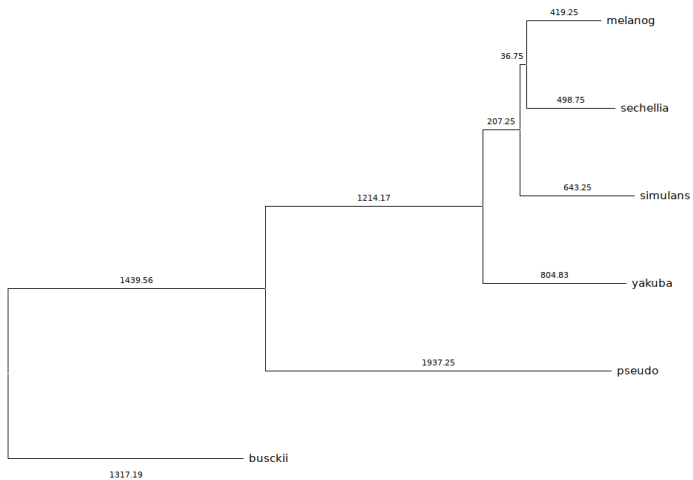
New: (1.4%) both genes are not part of any Flybase family;

Extension: (1.1%) one of the two genes is not part of any Flybase family;

Mismatch: (0.2%) each gene is in a different Flybase family.

Inferred phylogenies

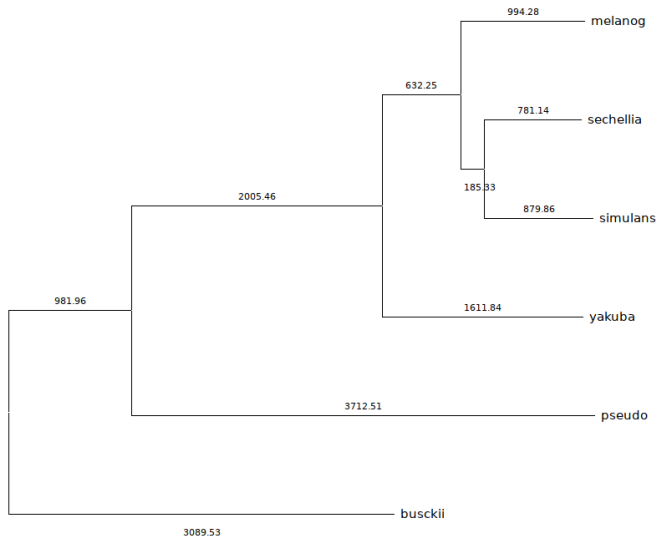
DING
with OMA families



500

Inferred phylogenies

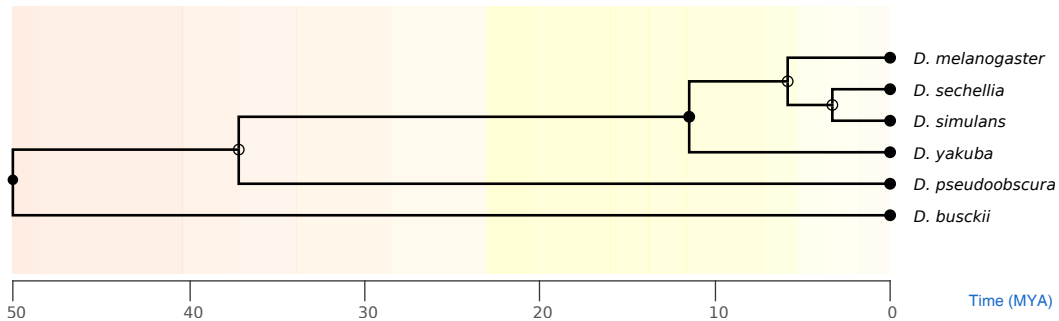
DIFF
with $x = 0.3$



1000

Reference phylogeny

TimeTree



References

On the family-free DCJ distance and similarity

(Fábio V. Martinez, Pedro Feijão, Marília D. V. Braga and Jens Stoye)

Algorithms for Molecular Biology (2015)

Natural family-free genomic distance

(Diego P. Rubert, Fábio V. Martinez and Marília D. V. Braga)

Proc. of WABI - LIPIcs, vol. 172 (2020)