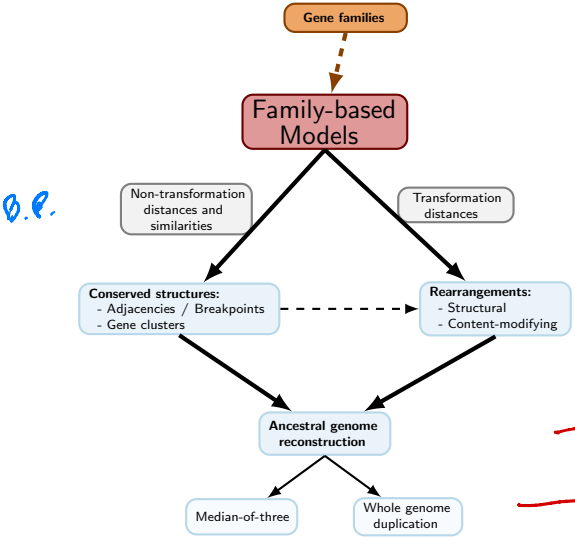
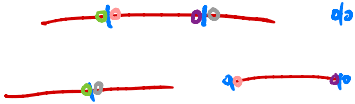
Topics of today:

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

Family-based setting



SCJ
INVERSION
DCJ



Are family assignments accurate?

Ideal situation:

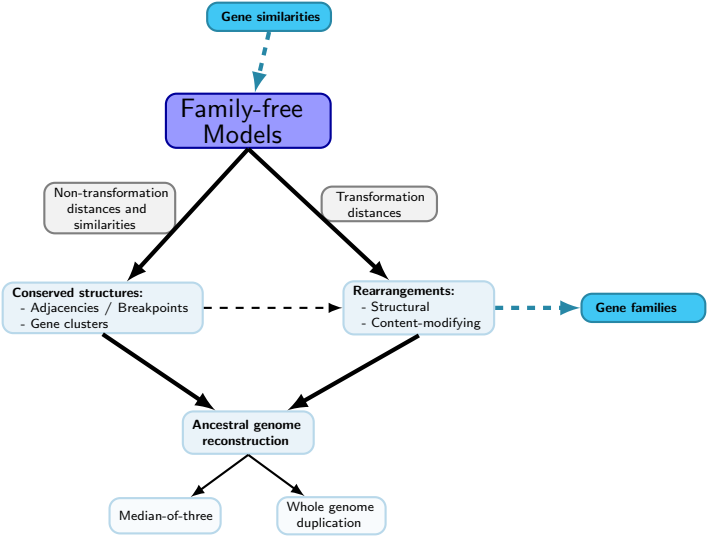
Are family assignments accurate?

Ideal situation:

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

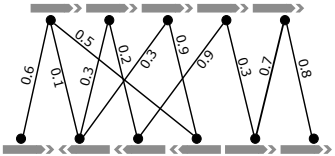
Reality:

Alternative: family-free setting



Family-free DCJ distance

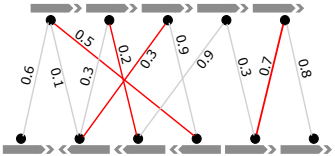
No family assignments , but pairwise normalized similarities
(above some threshold $x \in [0, 1]$)



$x = 0.1$

Family-free DCJ distance

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



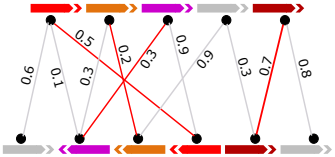
$x = 0.1$

$|M| = 4$
$w(M) = 1.7$

matching M

Family-free DCJ distance

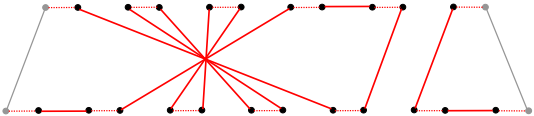
No family assignments , but pairwise normalized similarities
(above some threshold $x \in [0, 1]$)



$x = 0.1$

$|M| = 4$
$w(M) = 1.7$

matching M $\rightarrow$ singular mapped genomes $\mathbb{A}^M$ and $\mathbb{B}^M$
$\downarrow$
capped relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



capping of canonical genomes
(ignores indels/recombinations)

$d_{DCJ} = p_* + n - |K|$

$n = |M|$
$d_{DCJ} = 1 + 4 - 4 = 1$

Taking the weights into consideration

Weighted DCJ distance of mapped genomes

$$\mathrm{wd}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = \mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) \quad + \quad \boxed{|M| - w(M)}$$

$$= p_* + |M| - |\mathcal{C}| \quad + \quad |M| - w(M)$$

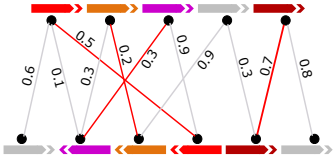
$$= p_* + 2|M| - |\mathcal{C}| - w(M)$$

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

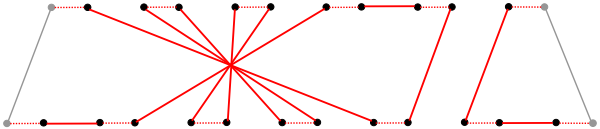
genes that are not covered by M are simply ignored

Weighted and unweighted DCJ distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$\mathrm{wd}_{\mathrm{DCJ}}$
M	4	1	2.3	3.3

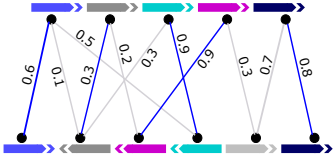


$$\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 4 - 4 = 1$$

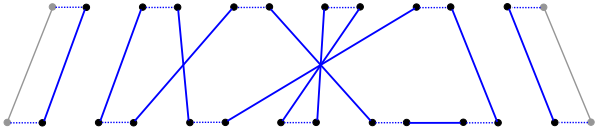
$$\mathrm{wd}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 2.3 = 3.3$$

Weighted and unweighted DCJ distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$\mathrm{wd}_{\mathrm{DCJ}}$
M	4	1	2.3	3.3
M	5	2	1.5	3.5

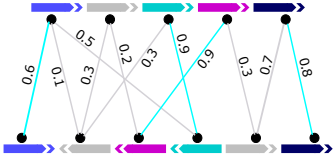


$$\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 5 - 4 = 2$$

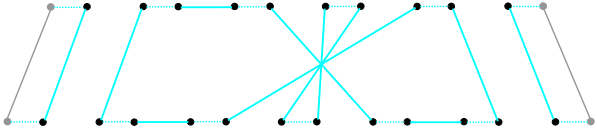
$$\mathrm{wd}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 2 + 1.5 = 3.5$$

Weighted and unweighted DCJ distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$\mathrm{wd}_{\mathrm{DCJ}}$
M	4	1	2.3	3.3
M	5	2	1.5	3.5
M	4	1	0.8	1.8

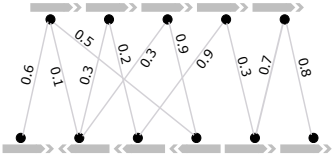


$$\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 4 - 4 = 1$$

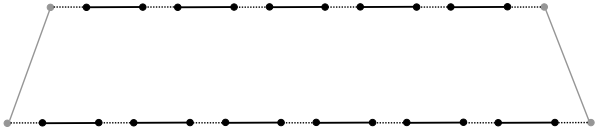
$$\mathrm{wd}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 0.8 = 1.8$$

Weighted and unweighted DCJ distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$\mathrm{wd}_{\mathrm{DCJ}}$
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

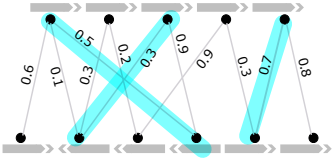


$$\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 0 - 1 = 0$$

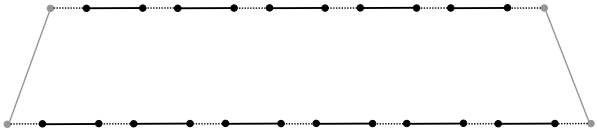
$$\mathrm{wd}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 0 + 0 = 0$$

Weighted and unweighted DCJ distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$\mathrm{wd}_{\mathrm{DCJ}}$
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~~
⋮	⋮	⋮	⋮	⋮



$$\mathrm{d}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 0 - 1 = 0$$

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

Family-free DCJ distance

$$\min_{M \in \mathfrak{M}_{\mathrm{MAX}}} \left\{ \mathrm{wd}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) \right\}$$

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

NP-hard

ILP formulation for the family-free DCJ distance

We have a capped multi-relational graph, but here each gene can be potentially ignored:

$\Rightarrow$ each gene has an indel edge and ignoring a gene is done by selecting its indel edge

Complete ILP formulation: extension of Shao-Lin-Moret, to be solved as an exercise

Quiz 1

1 In the FF DCJ formula the unmatched genes are...

A taken into consideration.

B simply ignored.

2 The weights in the FF DCJ formula penalizes...

A each pair of matched genes with similarity greater than 0.

B each pair of matched genes with similarity smaller than 1.

3 For computing the FF DCJ distance we need to select a sibling set (matching of genes)...

A of maximal size.

B of any size.

Job Announcement: Hilfskraftstelle

Tasks:

- fix a bug, test and release a new version of UNIMOG
  (java implementation of many rearrangement models for canonical/singular genomes)
  Available at `https://bibiserv.cebitec.uni-bielefeld.de/dcj`

- other small tasks concerning the inference of gene families via FF rearrangements

Workload: 10h/week

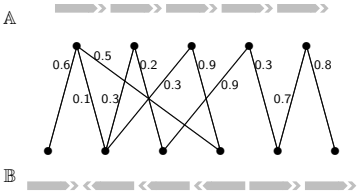
Duration: 4.5 months + 5 months

Remuneration (per month): SHK (without BSc) or WHK (with BSc)

Are you interested? Please contact me directly: mbraga@cebitec.uni-bielefeld.de

Family-free DCJ-indel distance

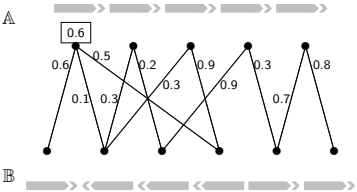
No family assignments , but pairwise normalized similarities (above some threshold $x \in [0, 1]$)



$x = 0.1$

Family-free DCJ-indel distance

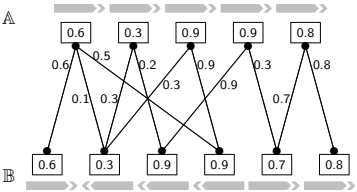
No family assignments , but pairwise normalized similarities (above some threshold $x \in [0, 1]$)



$x = 0.1$

Family-free DCJ-indel distance

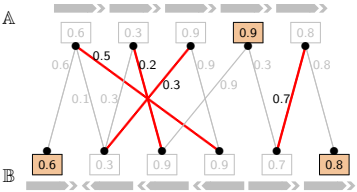
No family assignments , but pairwise normalized similarities (above some threshold $x \in [0, 1]$)



$x = 0.1$

Family-free DCJ-indel distance

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



matching M

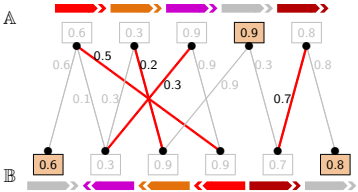
$x = 0.1$

$|M| = 4$

$w(M) = 1.7$

Family-free DCJ-indel distance

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



$x = 0.1$

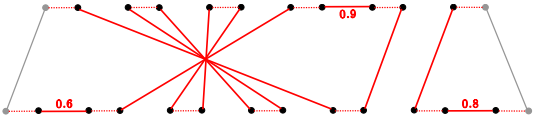
$|M| = 4$

$w(M) = 1.7$

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

$\downarrow$

weighted capped relational diagram



weight of indel edge = **maximum similarity** to the corresponding marker

capping of singular genomes (with indels/recombinations)

$\widetilde{M}$: set of indel edges (**complement** of M)

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

Taking the weights into consideration

Weighted DCJ-indel distance of mapped genomes

$$
\begin{aligned}
\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) &= \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) \quad + \quad \boxed{|M| - w(M)} + \boxed{w(\widetilde{M})} \\
&= p_* + |M| - |\mathcal{C}| + \textstyle\sum_{c \in \mathcal{C} \cup \mathcal{S}} \lambda(c) \quad + \quad |M| - w(M) + w(\widetilde{M}) \\
&= p_* + |M| - |\mathcal{C}| + |\mathcal{C}^r| + |\mathcal{S}| + \tfrac{\aleph}{2} \quad + \quad |M| - w(M) + w(\widetilde{M}) \\
&= p_* + |M| - |\mathcal{C}^{\bar{r}}| + |\mathcal{S}| + \tfrac{\aleph}{2} \quad + \quad |M| - w(M) + w(\widetilde{M})
\end{aligned}
$$

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

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

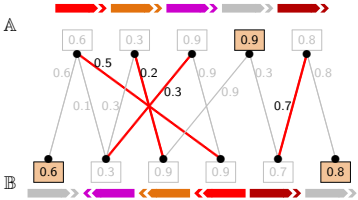
$\mathcal{C}^r$: set of indel-enclosing cycles

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

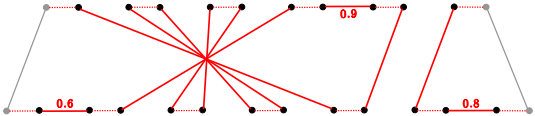
$\mathcal{S}$: set of circular singletons

Weighted and unweighted DCJ-indel distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$w(\widetilde{M})$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
M	4	4	2.3	2.3	8.6

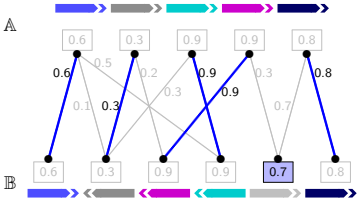


$$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 4 - 2 + 0 + \frac{2}{2} = 4$$

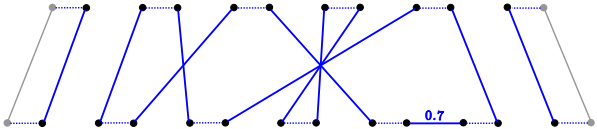
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 4 + 4 - 1.7 + 2.3 = 8.6$$

Weighted and unweighted DCJ-indel distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$w(\widetilde{M})$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
M	4	4	2.3	2.3	8.6
M	5	3	1.5	0.7	5.2

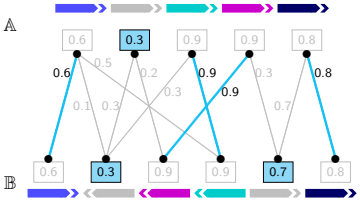


$$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 5 - 3 + 0 + 0 = 3$$

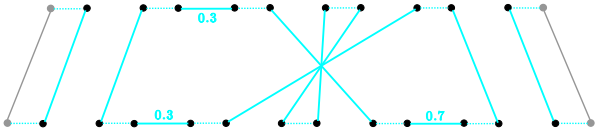
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 3 + 5 - 3.5 + 0.7 = 5.2$$

Weighted and unweighted DCJ-indel distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$w(\widetilde{M})$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
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

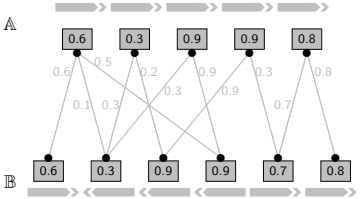


$$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 4 - 3 + 0 + \frac{2}{2} = 3$$

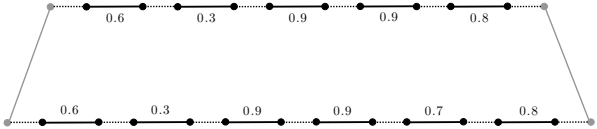
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 3 + 4 - 3.2 + 1.3 = 5.1$$

Weighted and unweighted DCJ-indel distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$w(\widetilde{M})$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
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
M	0	2	0	7.7	9.7

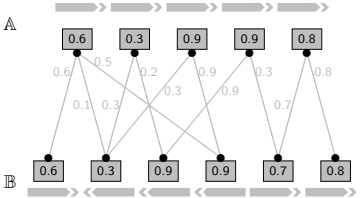


$$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 1 + 0 - 0 + 0 + \frac{2}{2} = 2$$

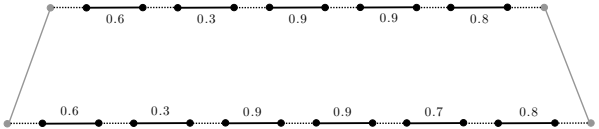
$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) = 2 + 0 - 0 + 7.7 = 9.7$$

Weighted and unweighted DCJ-indel distances of mapped genomes

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



M	$\lvert M\rvert$	$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}$	$\lvert M\rvert - w(M)$	$w(\widetilde{M})$	$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}$
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
M	0	2	0	7.7	9.7
⋮	⋮	⋮	⋮	⋮	⋮



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

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

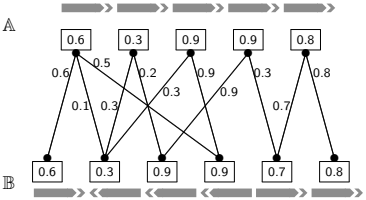
Family-free DCJ-indel distance

$$\min_{M \in \mathfrak{M}} \left\{ \mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A}^M, \mathbb{B}^M) \right\}$$

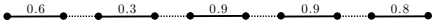
$\mathfrak{M}$: set of all matchings

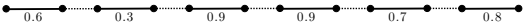
NP-hard

Approach for ILP



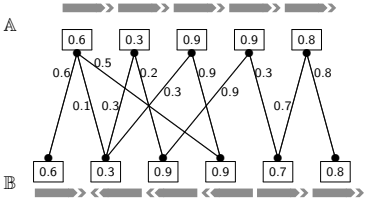
Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



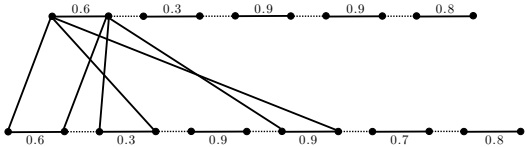


each marker has its **weighted indel edge** (weight = max. similarity)

Approach for ILP



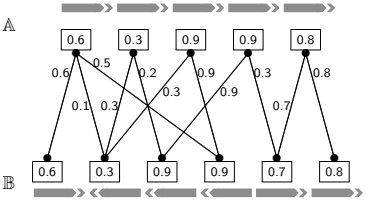
Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



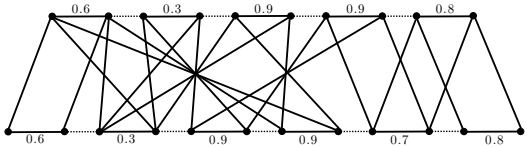
sibling weights are omitted

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

Approach for ILP



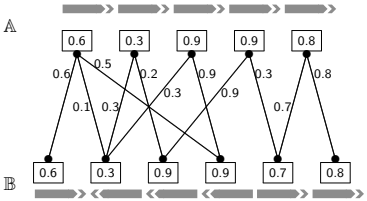
Capped family-free relational diagram : $p_* = \max\{\kappa(\mathbb{A}), \kappa(\mathbb{B})\}$



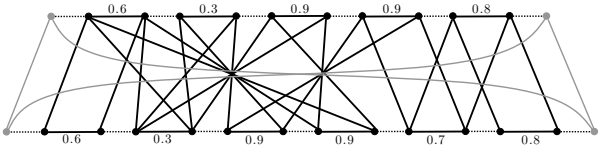
sibling weights are omitted

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

Approach for ILP



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



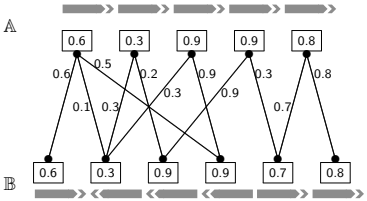
sibling weights are omitted

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

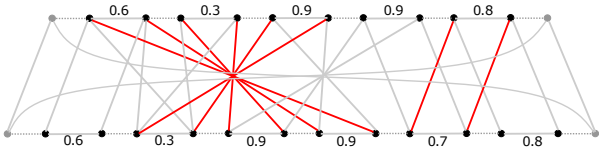
add $2p_*$ vertices (cap extremities)

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

Approach for ILP



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

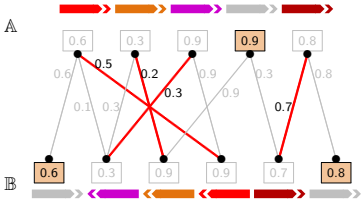


sibling weights are omitted

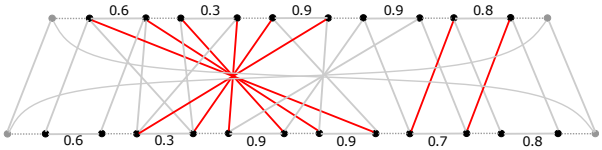
matching of extremity edges:

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

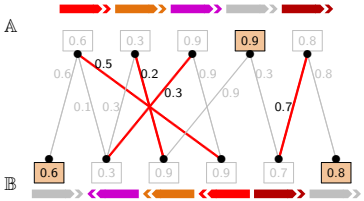
$|S| = 2|M|$

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

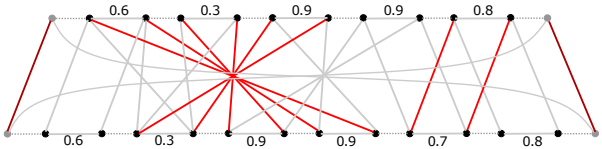
matching of extremity edges:

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

Approach for ILP



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



sibling weights are omitted

$|S| = 2|M|$

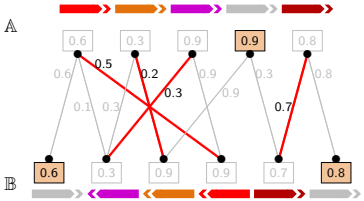
$w(S) = 2w(M)$

matching of extremity edges:

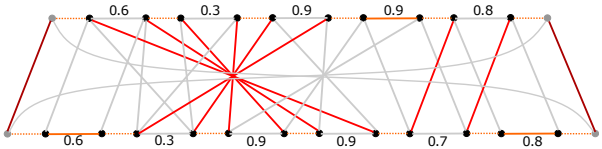
sibling set $S \to M$
(pairs of siblings)

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 $S \to M$
(pairs of siblings)

maximal capping set P
(covers all cap extremities)

capped consistent decomposition $Q[S, P]$

$$\begin{cases} S \cup P \\ \text{all adjacency edges} \\ \text{complement } \widetilde{S} \equiv \widetilde{M} \end{cases}$$

Optimization formula

DCJ-indel distance of a capped consistent decomposition

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

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

Optimization formula

DCJ-indel distance of a capped consistent decomposition

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

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

Weighted DCJ-indel distance of a capped consistent decomposition

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

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

Optimization formula

DCJ-indel distance of a capped consistent decomposition

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

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

Weighted DCJ-indel distance of a capped consistent decomposition

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

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

Family-free DCJ-indel distance

$$\min_{S \in \mathfrak{S}, P \in \mathfrak{P}_{\mathrm{MAX}}} \left\{ \mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(Q[S,P]) \right\}$$

$\mathfrak{S}$: set of sibling sets

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

Quiz 2

1 For computing the FF DCJ-indel distance we need to select a sibling set (matching of genes)...

A of maximal size.

B of any size.

2 The weights of the complement in the FF DCJ-indel formula penalizes...

A each unmatched gene with similarity to other genes greater than 0.

B each unmatched gene with similarity to other genes up to 0.5.

C each unmatched gene with similarity to other genes smaller than 1.

3 In the ILP formulation with the capped multi-relational graph, the path recombinations of the DCJ-indel distance...

A are simply ignored.

B are sometimes taken into consideration.

C are embedded in the capping.

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)

$$
\begin{aligned}
x_a &= 1 \quad \forall\, a \in E_\Gamma^{\mathbb{A}} \cup E_\Gamma^{\mathbb{B}} \\
\sum_{uv \in E} x_{uv} &= 2 \quad \forall\, u \in V \\
x_e &= x_d \quad \forall\, e, d \in E_\xi,\ e, d \text{ are siblings}
\end{aligned}
$$

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)

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

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

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

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)

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

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)

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

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

$$\sum_{\substack{d \in E_{\mathrm{ID}}^{\mathbb{A}} \\ d \cap e \neq \emptyset}} x_d - t_a \geq 0 \qquad \forall\ a \in E_{\mathrm{I}}^{\mathbb{A}}$$

$$t_e = 0 \qquad \forall\ e \in E \setminus E_{\mathrm{I}}^{\mathbb{A}}$$

ILP formulation for the family-free DCJ-indel distance

Weighted DCJ-indel distance formula

$$\mathrm{wd}^{\mathrm{ID}}_{\mathrm{DCJ}}(Q[S,P]) = p_* + |S| - |\mathcal{C}^{\tilde{r}}| + |\mathcal{S}| + \frac{\aleph}{2} - \frac{w(S)}{2} + w(\widetilde{S})$$

Objective function:

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

(one edge of the matching corresponds to a pair of edges in E_ξ)

Available at

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

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

ILP solver (alternative: GUROBI)

Pairwise comparisons of *Drosophila* genomes

$\sim$ 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	*pseudoobscura*	*sechellia*	*simulans*	*yakuba*	*busckii*
melanogaster	0.76%	4,431.78s	109.60s	201.49s	540.19s
pseudoobscura		163.12s	764.24s	5,782.73s	290.12s
sechellia			103.33s	146.88s	415.23s
simulans				216.77s	115.54s
yakuba					153.36s

(3h=10,800s):

Software for inferring gene families

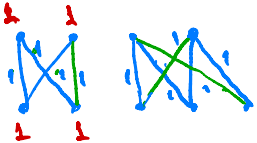
DING on OMA families with an average of 1.04 and at most 23 occurrences:

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



2. DING on families derived from similarity graphs extended cliques:
All but one comparisons reached the time limit of 3h.

1. 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: DING $\begin{cases} \text{has a smaller search space only composed of maximal sibling-sets} \\ \text{running times were considerably longer} \end{cases}$

Probable explanation:

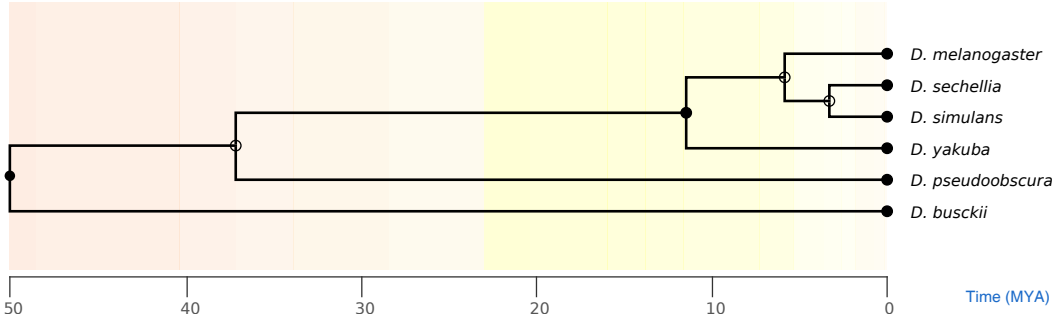
There is a larger number of co-optimal solutions in the DCJ-indel distance of natural genomes.
In the family-free DCJ-indel model the co-optimality is reduced by weights, allowing DIFF to converge faster.

⇒ 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.

3.

Reference phylogeny

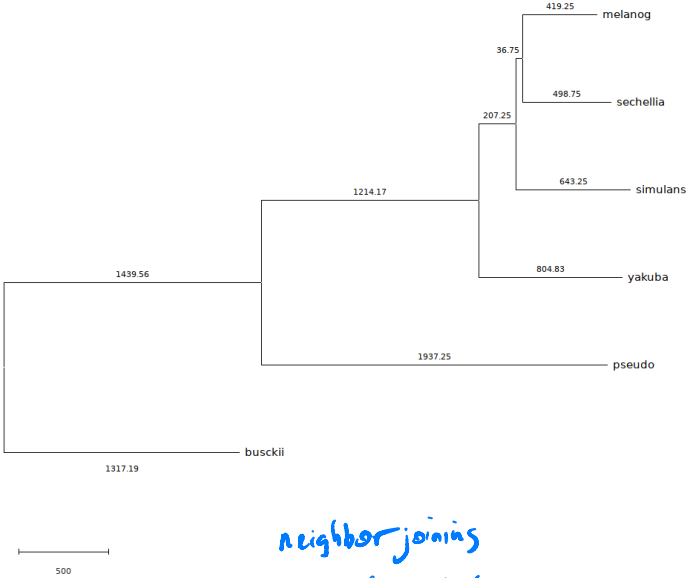
TimeTree



Inferred phylogenies

DING

with OMA families



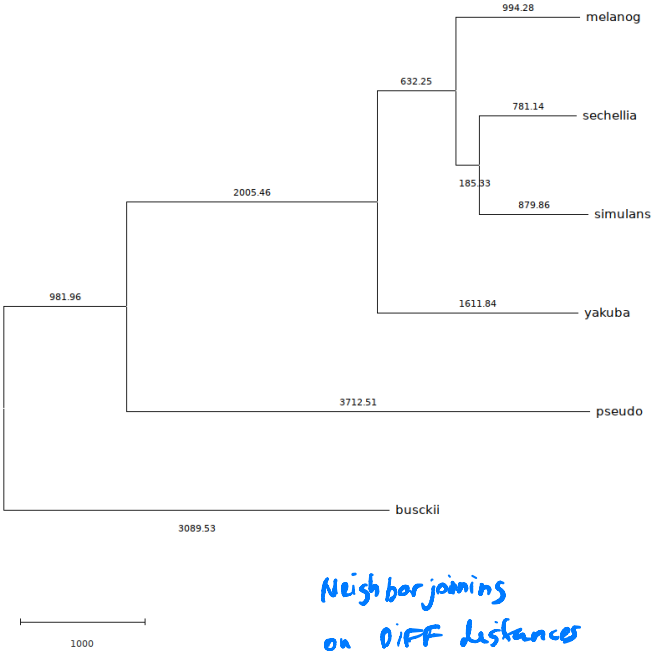
neighbor joining

on DING distances

Inferred phylogenies

DIFF

with $x = 0.3$



Neighbor joining on DIFF distances

Gene homologies established by DIFF compared to Flybase

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

$$\begin{cases} \textit{D. melanogaster} \\ \textit{D. simulans} \\ \textit{D. yakuba} \end{cases}$$

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.

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)

Algorithms for Molecular Biology (2021)