Topics of today:

More about singular DCJ-indel distance and sorting:

1. Restricted DCJ-indel model
2. The diameter of the DCJ-indel distance
3. Establishing the triangular inequality

Capping

1. Capped relational graph of canonical genomes
2. Capped relational graph of singular genomes
3. Indel-potential of cycles via transitions

Singular DCJ-indel model - summary

DCJ-indel distance: $d_{DCJ}^{ID}(\mathbb{A}, \mathbb{B}) = n - |\mathcal{C}| - \frac{|\mathcal{P}_{\mathbb{AB}}|}{2} + \sum_{C \in RG} \lambda(C) - \delta,$ where δ is the value obtained by optimizing deducting path recombinations

$\mathbb{A}$ and $\mathbb{B}$ are circular: $d_{DCJ}^{ID}(\mathbb{A}, \mathbb{B}) = n - |\mathcal{C}| + \sum_{C \in RG} \lambda(C)$

Both distance computation and sorting can be done in **linear time.**

Singular DCJ-indel sorting: trade-off between DCJ and indels

A sorting algorithm that maximizes gaining DCJs with $\Delta_\lambda = 0$ ~~minimizes~~ maximizes indels.

However, these gaining DCJs can be often replaced by $\begin{cases} \text{neutral DCJs with } \Delta_\lambda = -1 \\ \text{losing DCJs with } \Delta_\lambda = -2 \end{cases}$

$\Downarrow$

There is a big range of sorting possibilities between

$$\begin{cases} \text{a sorting algorithm that maximizes gaining DCJs with } \Delta_\lambda = 0 \text{ (minimizing indels)} \\ \text{a sorting algorithm that minimizes gaining DCJs with } \Delta_\lambda = 0 \text{ (maximizing indels)} \end{cases}$$

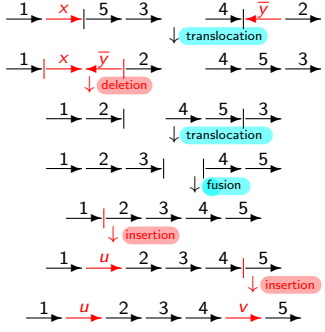
Singular DCJ-indel sorting: trade-off between DCJ and indels

Two optimal scenarios sorting $\{\ [1\,x\,5\,3],\ [4\,\overline{y}\,2]\ \}$ into $\{\ [1\,u\,2\,3\,4\,v\,5]\ \}$

Maximizing gaining DCJs

(i)

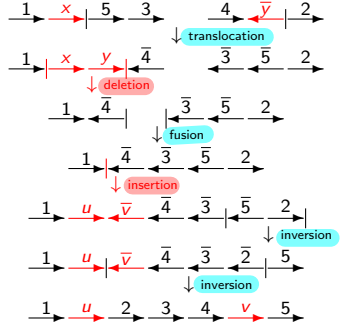
Minimizing DCJs gives 3 DCJs and 3 indels



Minimizing gaining DCJs

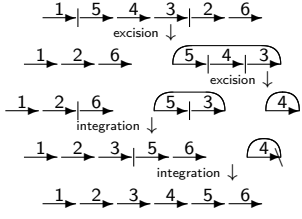
(ii)

Minimizing indels gives 4 DCJs and 2 indels



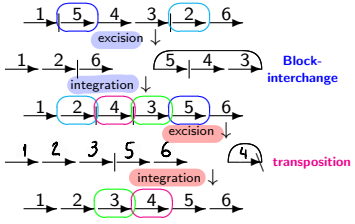
DCJ model - circular excision/integration

Canonical DCJ model



Many circular chromosomes can coexist in the intermediate genomes.

Restricted canonical DCJ model



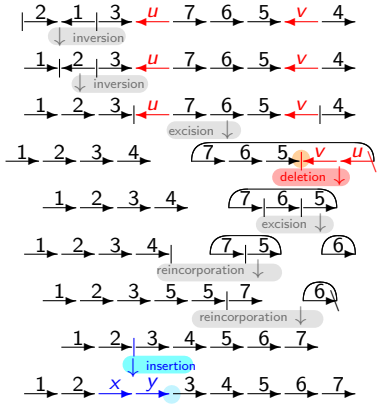
A circular chromosome is immediately reintegrated after its excision.

The DCJ distance is the same
for both the general and the restricted DCJ models

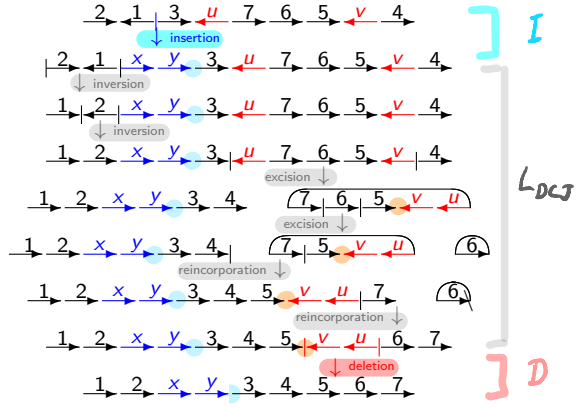
Restricted DCJ-indel (singular linear genomes)

In any sorting sequence, it is always possible to $\begin{cases} \text{move } \textbf{deletions down} \\ \text{move } \textbf{insertions up} \end{cases}$

S: general DCJ-indel sorting



L: "layered" DCJ-indel sorting

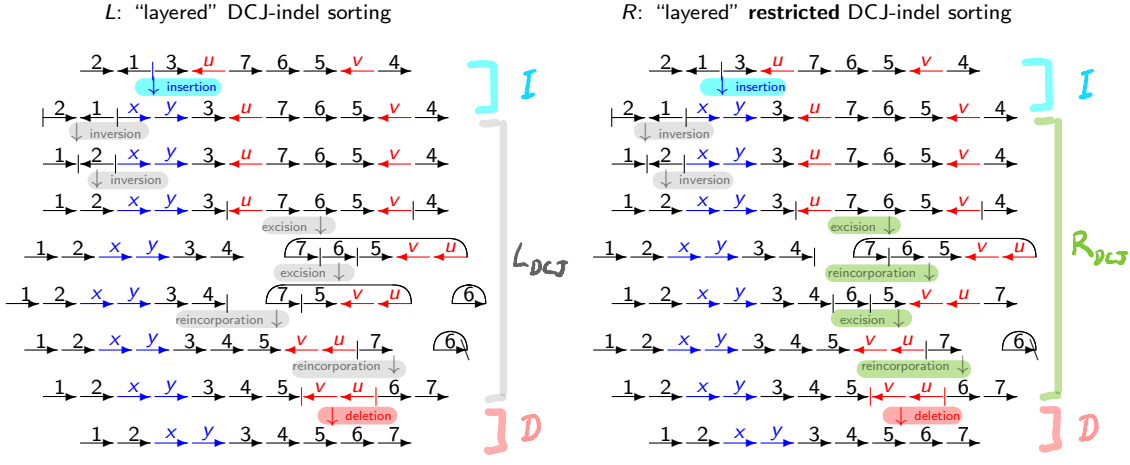


S : general sequence of DCJ and indel operations sorting linear $\mathbb{A}$ into linear $\mathbb{B}$

L : "layered" sequence of insertions, DCJs and deletions sorting linear $\mathbb{A}$ into linear $\mathbb{B}$

$$S \quad \rightsquigarrow \quad L = I \oplus L_{\mathrm{DCJ}} \oplus D \qquad \text{and} \quad |S| = |L|$$

Restricted DCJ-indel-distance (singular linear genomes)



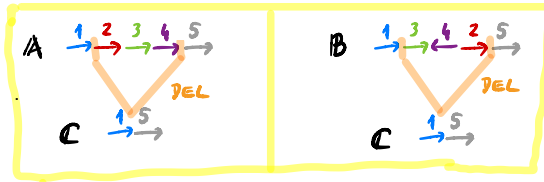
L : "layered" sequence of insertions, DCJs and deletions sorting linear $\mathbb{A}$ into linear $\mathbb{B}$

R : "layered" sequence of insertions, restricted DCJs and deletions sorting linear $\mathbb{A}$ into linear $\mathbb{B}$

$$S \quad \rightsquigarrow \quad L = I \oplus L_{\mathrm{DCJ}} \oplus D \quad \rightsquigarrow \quad R = I \oplus R_{\mathrm{DCJ}} \oplus D \qquad \text{and} \quad |S| = |S'| = |R|$$

The triangular inequality does not hold for the DCJ-indel distance

Three singular genomes $\begin{cases} \mathbb{A} = [1\ 2\ 3\ 4\ 5] \\ \mathbb{B} = [1\ 3\ \overline{4}\ 2\ 5] \\ \mathbb{C} = [1\ 5] \end{cases}$



The triangular inequality

$d^{ID}_{DCJ}(\mathbb{A}, \mathbb{B}) \leq d^{ID}_{DCJ}(\mathbb{A}, \mathbb{C}) + d^{ID}_{DCJ}(\mathbb{B}, \mathbb{C})$

does not hold

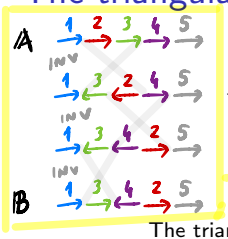
$\begin{cases} d^{ID}_{DCJ}(\mathbb{A}, \mathbb{B}) = 3 \\ d^{ID}_{DCJ}(\mathbb{A}, \mathbb{C}) = 1 \\ d^{ID}_{DCJ}(\mathbb{B}, \mathbb{C}) = 1 \end{cases}$

"Free lunch":
while sorting $\mathbb{A}$ into $\mathbb{C}$ and then $\mathbb{C}$ into $\mathbb{B}$,
a set of common genes of $\mathbb{A}$ and $\mathbb{B}$
are deleted and then reinserted

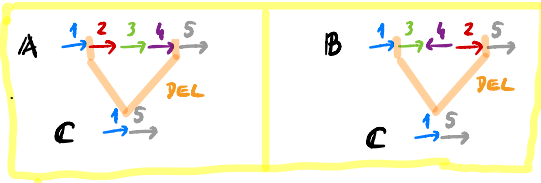
It is possible to define a "corrected" DCJ-indel distance, for which the triangular inequality is fulfilled.

The first step towards this correction is determining the DCJ-indel diameter

The triangular inequality does not hold for the DCJ-indel distance



Three singular genomes $\begin{cases} \mathbb{A} = [\,1\ 2\ 3\ 4\ 5\,] \\ \mathbb{B} = [\,1\ 3\ \overline{4}\ 2\ 5\,] \\ \mathbb{C} = [\,1\ 5\,] \end{cases}$



The triangular inequality

$d^{ID}_{DCJ}(\mathbb{A}, \mathbb{B}) \leq d^{ID}_{DCJ}(\mathbb{A}, \mathbb{C}) + d^{ID}_{DCJ}(\mathbb{B}, \mathbb{C})$

does not hold

$\begin{cases} d^{ID}_{DCJ}(\mathbb{A}, \mathbb{B}) = 3 \\ d^{ID}_{DCJ}(\mathbb{A}, \mathbb{C}) = 1 \\ d^{ID}_{DCJ}(\mathbb{B}, \mathbb{C}) = 1 \end{cases}$

"Free lunch":
while sorting $\mathbb{A}$ into $\mathbb{C}$ and then $\mathbb{C}$ into $\mathbb{B}$,
a set of common genes of $\mathbb{A}$ and $\mathbb{B}$
are deleted and then reinserted

It is possible to define a "corrected" DCJ-indel distance, for which the triangular inequality is fulfilled.

The first step towards this correction is determining the DCJ-indel diameter

OBS: In the comparison of two genomes, the DCJ-indel model prevents the "free lunch":
common genes cannot be deleted or inserted

The diameter D_{DCJ}^{ID} of the DCJ-indel-distance

The **diameter** is the tightest known upper bound for a **distance measure**.

Denote by $D_{DCJ}^{ID}(\mathbb{A}, \mathbb{B})$ the DCJ-indel diameter of two singular genomes $\mathbb{A}$ and $\mathbb{B}$:

$$d_{DCJ}^{ID}(\mathbb{A}, \mathbb{B}) \quad \leq \quad D_{DCJ}^{ID}(\mathbb{A}, \mathbb{B})$$

For determining the diameter $D_{DCJ}^{ID}(\mathbb{A}, \mathbb{B})$, we

- ignore path recombinations and consider that the components of $RG(\mathbb{A}, \mathbb{B})$ are sorted independently
- determine the maximum number of runs that each component of $RG(\mathbb{A}, \mathbb{B})$ could have

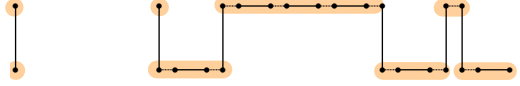
For a given component C in a relational graph, let a **segment** of C be

- C itself (if C is a 0-cycle or a 0-path)
- a minimal path flanked by two extremity-edges
- a minimal path at the extremity of a path and connected to an extremity edge

$\beth(C)$: number of segments in component C

$\beth(C)$ depends on

- the type of C (path or cycle)
- the number of extremity edges in C



$\beth(p_1) = 2$ $\beth(p_5) = 6$

Segments of a component in the relational graph

For a given component C in a relational graph, let a **segment** of C be

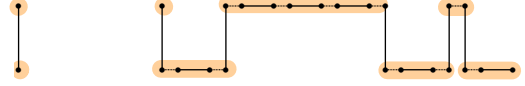
$$\begin{cases} C \text{ itself} & \text{(if } C \text{ is a 0-cycle or a 0-path)} \\ \text{a minimal path flanked by two extremity-edges} \\ \text{a minimal path at the extremity of a path and connected to an extremity edge} \end{cases}$$

$\beth(C)$: number of segments in component C

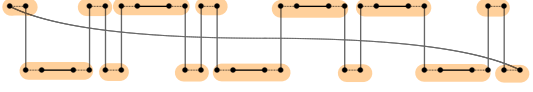
$\beth(C)$ depends on

$$\begin{cases} \text{the type of } C \text{ (path or cycle)} \\ \text{the number of extremity edges in } C \end{cases}$$

in general $\begin{cases} \beth(p_k) = k+1 & k \geq 0 \\ \beth(c_0) = 1 \\ \beth(c_k) = k & k \in \{2, 4, 6, ...\} \end{cases}$



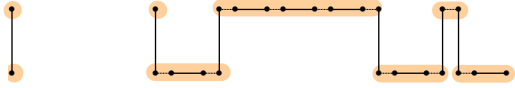
$\beth(p_1) = 2$ $\beth(p_5) = 6$



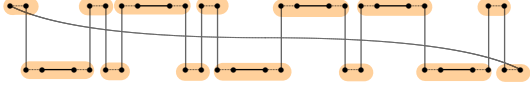
$\beth(c_{14}) = 14$

Segments of a component in the relational graph

$\beth(C)$: number of segments in component C



$\beth(p_1) = 2$ $\qquad$ $\beth(p_5) = 6$



$\beth(c_{14}) = 14$

in general $\begin{cases} \beth(p_k) = k+1 & k \geq 0 \\ \beth(c_0) = 1 & \\ \beth(c_k) = k & k \in \{2, 4, 6, \ldots\} \end{cases}$

$\Lambda_{\text{MAX}}(C)$: max. number of runs that C can have

$\lambda_{\text{MAX}}(C)$: max. indel-potential that C can have

$\Lambda_{\text{MAX}}(C) = \beth(C)$ and $\lambda_{\text{MAX}}(C) = \left\lceil \frac{\beth(C)+1}{2} \right\rceil$

$\beth(C)$	$\Lambda_{\text{MAX}}(C)$	$\lambda_{\text{MAX}}(C)$	$\text{d}_{\text{DCJ}}(C)$
1	1	1	0
2	2	2	0
3	3	2	1
4	4	3	1
5	5	3	2
6	6	4	2
7	7	4	3
⋮	⋮	⋮	⋮
$\beth(C)$	$\beth(C)$	$\left\lceil \frac{\beth(C)+1}{2} \right\rceil$	$\left\lfloor \frac{\beth(C)-1}{2} \right\rfloor$

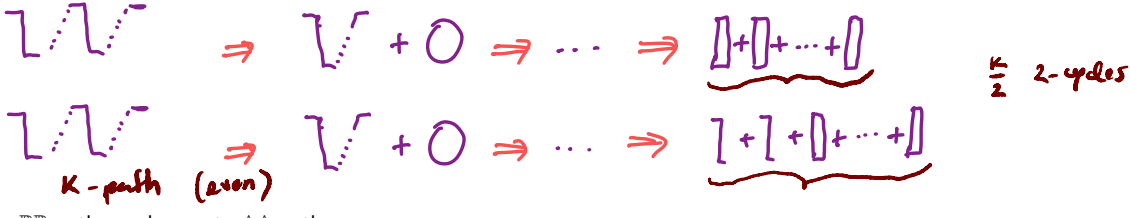
Each component can be sorted separately...

...with an internal gaining DCJ at each step:

Cycle: creates a new cycle at each step

$\mathbb{AB}$-path: creates a new cycle at each step

$\mathbb{AA}$-path: creates a new cycle at each step, eventually one step is a single cut (on $\mathbb{B}$) that creates two $\mathbb{AB}$-paths

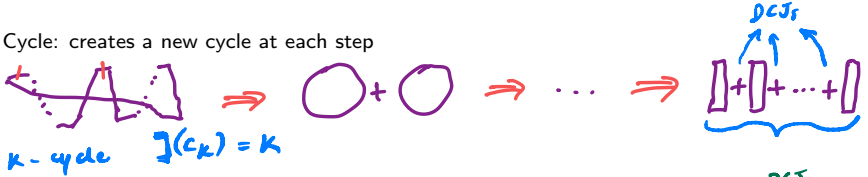


$\mathbb{BB}$-path: analogous to $\mathbb{AA}$-path

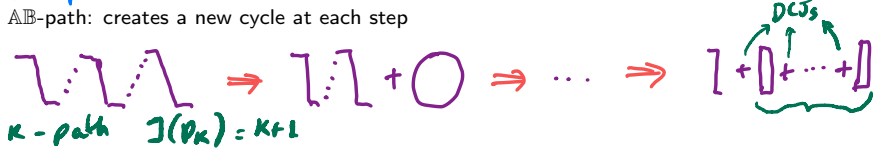
Each component can be sorted separately...

...with an internal gaining DCJ at each step:

Cycle: creates a new cycle at each step



$\mathbb{AB}$-path: creates a new cycle at each step



$\mathbb{AA}$-path: creates a new cycle at each step, eventually one step is a single cut (on $\mathbb{B}$) that creates two $\mathbb{AB}$-paths

$\mathbb{BB}$-path: analogous to $\mathbb{AA}$-path

$$d_{DCJ}(C) = \left\lceil \frac{J(C) - 1}{2} \right\rceil$$

The diameter D_{DCJ}^{ID} of the DCJ-indel-distance

$\beth(C)$: number of segments in component C

$\beth(C)$	$\Lambda_{MAX}(C)$	$\lambda_{MAX}(C)$	$d_{DCJ}(C)$
1	1	1	0
2	2	2	0
3	3	2	1
4	4	3	1
5	5	3	2
6	6	4	2
7	7	4	3
⋮	⋮	⋮	⋮
$\beth(C)$	$\beth(C)$	$\left\lceil \frac{\beth(C)+1}{2} \right\rceil$	$\left\lfloor \frac{\beth(C)-1}{2} \right\rfloor$

if $\beth(C)$ is odd:

$$d_{DCJ}(C) + \lambda_{MAX}(C) = \frac{\beth(C)-1}{2} + \frac{\beth(C)+1}{2} = \beth(C)$$

if $\beth(C)$ is even:

$$d_{DCJ}(C) + \lambda_{MAX}(C) = \frac{\beth(C)-2}{2} + \frac{\beth(C)+2}{2} = \beth(C)$$

Let $\begin{cases} \kappa(\mathbb{A}): & \#\text{ linear chromosomes in } \mathbb{A} \\ \mathcal{S}(\mathbb{A}): & \#\text{ (circular) singletons in } \mathbb{A} \\ \kappa(\mathbb{B}): & \#\text{ linear chromosomes in } \mathbb{B} \\ \mathcal{S}(\mathbb{B}): & \#\text{ (circular) singletons in } \mathbb{B} \end{cases}$

The number of segments in $RG(\mathbb{A}, \mathbb{B})$ is

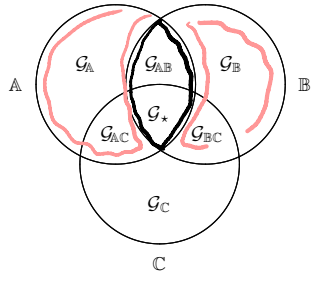
$$\beth(RG(\mathbb{A}, \mathbb{B})) = 2n + \kappa(\mathbb{A}) + \mathcal{S}(\mathbb{A}) + \kappa(\mathbb{B}) + \mathcal{S}(\mathbb{B})$$

$$\begin{aligned} D_{DCJ}^{ID}(\mathbb{A}, \mathbb{B}) &= \sum_{C \in RG(\mathbb{A},\mathbb{B})} (d_{DCJ}(C) + \lambda_{MAX}(C)) \\ &= \sum_{C \in RG(\mathbb{A},\mathbb{B})} \beth(C) \\ &= \beth(RG(\mathbb{A}, \mathbb{B})) \end{aligned}$$

$$\boxed{D_{DCJ}^{ID}(\mathbb{A}, \mathbb{B}) = 2n + \kappa(\mathbb{A}) + \mathcal{S}(\mathbb{A}) + \kappa(\mathbb{B}) + \mathcal{S}(\mathbb{B})}$$

Establishing the triangular inequality

Disjoint sets of genes $\mathcal{G}_{\mathbb{A}}$, $\mathcal{G}_{\mathbb{B}}$, $\mathcal{G}_{\mathbb{C}}$, $\mathcal{G}_{\mathbb{AB}}$, $\mathcal{G}_{\mathbb{BC}}$, $\mathcal{G}_{\mathbb{AC}}$ and $\mathcal{G}_{\star}$
for three genomes $\mathbb{A}, \mathbb{B}$ and $\mathbb{C}$



For each pair of genomes, we define the **corrected distance** $\mathsf{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}$:

$$\mathsf{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) = \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) + \mathsf{k}(|\mathcal{G}_{\mathbb{A}}| + |\mathcal{G}_{\mathbb{AC}}| + |\mathcal{G}_{\mathbb{B}}| + |\mathcal{G}_{\mathbb{BC}}|)$$
$$\mathsf{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) = \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + \mathsf{k}(|\mathcal{G}_{\mathbb{A}}| + |\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}| + |\mathcal{G}_{\mathbb{BC}}|)$$
$$\mathsf{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) = \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) + \mathsf{k}(|\mathcal{G}_{\mathbb{B}}| + |\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}| + |\mathcal{G}_{\mathbb{AC}}|)$$

The triangular inequality must hold for $\mathsf{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}$:

$$\mathsf{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) \leq \mathsf{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + \mathsf{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C})$$

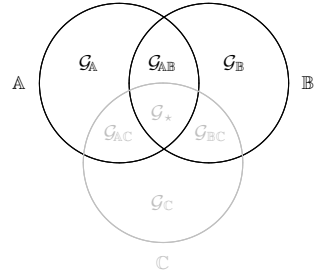
$$\begin{aligned}\mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) + \mathsf{k}(|\mathcal{G}_{\mathbb{A}}| + |\mathcal{G}_{\mathbb{AC}}| + |\mathcal{G}_{\mathbb{B}}| + |\mathcal{G}_{\mathbb{BC}}|) &\leq \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + \mathsf{k}(|\mathcal{G}_{\mathbb{A}}| + |\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}| + |\mathcal{G}_{\mathbb{BC}}|) + \\ &\quad \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) + \mathsf{k}(|\mathcal{G}_{\mathbb{B}}| + |\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}| + |\mathcal{G}_{\mathbb{AC}}|)\end{aligned}$$

$$\mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) \leq \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + \mathsf{k}(|\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}|) + \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) + \mathsf{k}(|\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}|)$$

$$\mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) \leq \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + \mathsf{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) + 2\mathsf{k}(|\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}|)$$

Establishing the triangular inequality

$$\begin{cases} \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) \;\leq\; \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) + 2\mathsf{k}(|\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}|) \\ \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) \;\leq\; \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) + \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) + 2\mathsf{k}(|\mathcal{G}_{\mathbb{AC}}| + |\mathcal{G}_{\mathbb{B}}|) \\ \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) \;\leq\; \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) + \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + 2\mathsf{k}(|\mathcal{G}_{\mathbb{BC}}| + |\mathcal{G}_{\mathbb{A}}|) \end{cases}$$



Assume $\begin{cases} \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) \geq \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) \\ \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) \geq \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) \end{cases}$

Let $\begin{cases} \chi(\mathbb{A}): & \#\text{ chromosomes in } \mathbb{A} \\ \kappa(\mathbb{A}): & \#\text{ linear chromosomes in } \mathbb{A} \\ \mathcal{S}(\mathbb{A}): & \#\text{ (circular) singletons in } \mathbb{A} \\ \chi(\mathbb{B}): & \#\text{ chromosomes in } \mathbb{B} \\ \kappa(\mathbb{B}): & \#\text{ linear chromosomes in } \mathbb{B} \\ \mathcal{S}(\mathbb{B}): & \#\text{ (circular) singletons in } \mathbb{B} \end{cases}$

$\kappa(\mathbb{A}) + \mathcal{S}(\mathbb{A}) \leq \chi(\mathbb{A})$

and

$\kappa(\mathbb{B}) + \mathcal{S}(\mathbb{B}) \leq \chi(\mathbb{B})$

We need to find a value k that guarantees:

$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) \;\leq\; \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) + 2\mathsf{k}(|\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}|)$

In the worst case genome $\mathbb{C}$ is empty:

$\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) = \chi(\mathbb{A})$ and $\mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) = \chi(\mathbb{B})$

$\mathrm{D}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) = 2|\mathcal{G}_{\mathbb{AB}}| + \kappa(\mathbb{A}) + \mathcal{S}(\mathbb{A}) + \kappa(\mathbb{B}) + \mathcal{S}(\mathbb{B})$

$\mathrm{D}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) \leq \chi(\mathbb{A}) + \chi(\mathbb{B}) + 2\mathsf{k}|\mathcal{G}_{\mathbb{AB}}|$

$\vdots$

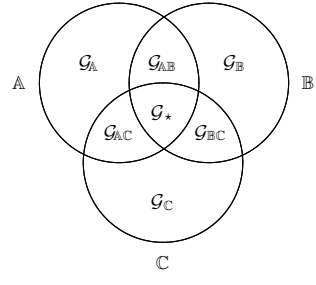
$2|\mathcal{G}_{\mathbb{AB}}| \;\leq\; 2\mathsf{k}|\mathcal{G}_{\mathbb{AB}}| \;\Rightarrow\; \boxed{\mathsf{k} \geq 1}$

Establishing the triangular inequality

$$\mathrm{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{B}) + \mathrm{k}(|\mathcal{G}_{\mathbb{A}}| + |\mathcal{G}_{\mathbb{AC}}| + |\mathcal{G}_{\mathbb{B}}| + |\mathcal{G}_{\mathbb{BC}}|)$$

$$\mathrm{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{A},\mathbb{C}) + \mathrm{k}(|\mathcal{G}_{\mathbb{A}}| + |\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}| + |\mathcal{G}_{\mathbb{BC}}|)$$

$$\mathrm{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) = \mathrm{d}^{\mathrm{ID}}_{\mathrm{DCJ}}(\mathbb{B},\mathbb{C}) + \mathrm{k}(|\mathcal{G}_{\mathbb{B}}| + |\mathcal{G}_{\mathbb{AB}}| + |\mathcal{G}_{\mathbb{C}}| + |\mathcal{G}_{\mathbb{AC}}|)$$



The triangular inequality holds for the corrected distance $\mathrm{dk}^{\mathrm{ID}}_{\mathrm{DCJ}}$ for any $\mathrm{k} \geq 1$

Quiz 1

1 Which of the following statements about the DCJ-indel model are true?

A The triangular inequality holds for the DCJ-indel distance.

B The triangular inequality does not hold for the DCJ-indel distance, but a simple correction can be done.

C The DCJ-indel distance can be distinct from the restricted DCJ-indel distance.

2 The best known algorithm for the restricted DCJ-indel sorting runs in...

A $O(n)$ time.

B $O(n \log n)$ time.

C $O(n^2)$ time.

Capped relational graph

Capping is a procedure that circularizes all paths of a relational graph by adding **caps** (**artificial genes**):

- if the capping is optimal, the genomic distance is preserved
- from the capped relational diagram we can derive genomes composed only of circular chromosomes

A capping may require adjacencies between caps:

$\Gamma_{\mathbb{A}}$: represents an adjacency between caps in genome $\mathbb{A}$

$\Gamma_{\mathbb{B}}$: represents an adjacency between caps in genome $\mathbb{B}$.

Capped relational graph of canonical genomes

Optimally linking paths from $RG(\mathbb{A}, \mathbb{B})$ of canonical genomes $\mathbb{A}$ and $\mathbb{B}$ into cycles can be done as follows:

id	paths	linking cycle		Δn	Δc	$\Delta(2\mathbb{AB})$	Δ_{DCJ}
1	$\mathbb{AB}$	$(\mathbb{AB})$		$+0.5$	$+1$	-0.5	0
2	$\mathbb{AA} + \mathbb{BB}$	$(\mathbb{AA}, \mathbb{BB})$		$+1$	$+1$	0	0
3	$\mathbb{AA}$	$(\mathbb{AA}, \Gamma_{\mathbb{B}})$	$\cup$	$+1$	$+1$	0	0
4	$\mathbb{BB}$	$(\mathbb{BB}, \Gamma_{\mathbb{A}})$	$\cap$	$+1$	$+1$	0	0

Closing an $\mathbb{AA}$-path (over-represented in genome $\mathbb{A}$ and marked with a $\cup$) requires an adjacency $\Gamma_{\mathbb{B}}$.

Closing a $\mathbb{BB}$-path (over-represented in genome $\mathbb{B}$ and marked with a $\cap$) requires an adjacency $\Gamma_{\mathbb{A}}$.

Any capping producing linking cycles as indicated on the table above is optimal:

- The value $\Delta_{\text{DCJ}} = \Delta n - \Delta c - \Delta(2\mathbb{AB})$ is the DCJ-effect produced by each type of linking cycle.
- All given linking cycles have $\Delta_{\text{DCJ}} = 0$, therefore they preserve the DCJ distance.

Let $\kappa_{\mathbb{A}}$: number of linear chromosomes in $\mathbb{A}$

$\kappa_{\mathbb{B}}$: number of linear chromosomes in $\mathbb{B}$

The difference between the number of $\mathbb{AA}$- and of $\mathbb{BB}$-paths is equal to the difference between $\kappa_{\mathbb{A}}$ and $\kappa_{\mathbb{B}}$.

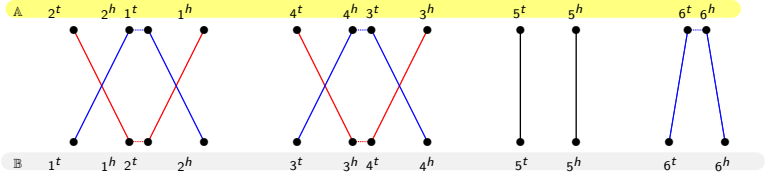
An optimal capping that maximizes the number of linking cycles of type **2** minimizes the number of caps:

The number of caps to be added is exactly $p_* = \max\{\kappa_{\mathbb{A}}, \kappa_{\mathbb{B}}\}$.

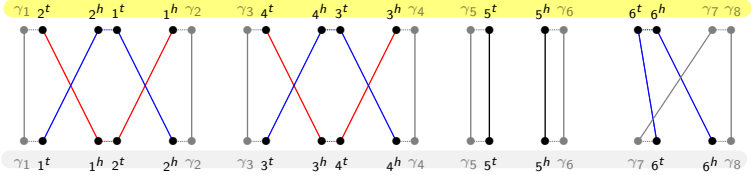
The number of adjacencies between caps is exactly $a_* = |\kappa_{\mathbb{A}} - \kappa_{\mathbb{B}}|$.

Capped relational graph of canonical genomes - example

$\mathbb{A} = [2\ 1]\ [4\ 3]\ [5]\ (6)$ and $\mathbb{B} = [1\ 2]\ [3\ 4]\ [5]\ [6]$; $p_* = 4$ and $a_* = 1$



$$\begin{aligned} \mathrm{d}_{\mathrm{DCJ}} &= n - |\mathcal{C}| - \frac{|\mathcal{P}_{\mathbb{AB}}|}{2} \\ &= 6 - 0 - 1 \\ &= 5 \end{aligned}$$



$$\begin{aligned} \mathrm{d}_{\mathrm{DCJ}} &= n + p_* - |\mathcal{C}| \\ &= 6 + 4 - 5 \\ &= 5 \end{aligned}$$

Any way of pairing the cap extremities $\gamma_1, \gamma_2, ..., \gamma_8$ is valid; possible derived circular genomes are:

$\mathbb{A}_\circ = (2\ 1\ \mathtt{W})\ (4\ 3\ \mathtt{X})\ (5\ \mathtt{Y})\ (6)\ (\mathtt{Z})$ and $\mathbb{B}_\circ = (1\ 2\ \mathtt{W})\ (3\ 4\ \mathtt{X})\ (5\ \mathtt{Y})\ (6\ \mathtt{Z})$

$(\mathtt{W}^h = \gamma_1, \mathtt{W}^t = \gamma_2, \mathtt{X}^h = \gamma_3, \mathtt{X}^t = \gamma_4, \mathtt{Y}^h = \gamma_5, \mathtt{Y}^t = \gamma_6, \mathtt{Z}^h = \gamma_7, \mathtt{Z}^t = \gamma_8)$

or

$\mathbb{A}_\circ = (2\ 1\ \mathtt{W}\ 4\ 3\ \mathtt{X}\ 5\ \ \mathtt{Y}\ \mathtt{Z})\ (6)$ and $\mathbb{B}_\circ = (1\ 2\ \mathtt{W}\ 3\ 4\ \mathtt{X}\ 5\ \mathtt{Y}\ 6\ \mathtt{Z})$

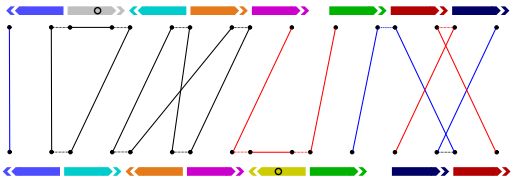
$(\mathtt{W}^h = \gamma_3, \mathtt{W}^t = \gamma_2, \mathtt{X}^h = \gamma_5, \mathtt{X}^t = \gamma_4, \mathtt{Y}^h = \gamma_7, \mathtt{Y}^t = \gamma_6, \mathtt{Z}^h = \gamma_1, \mathtt{Z}^t = \gamma_8)$

Capped relational graph

$n = \#$ common families

$p_* = \max \begin{cases} \#\text{ linear chromosomes in } A \\ \#\text{ linear chromosomes in } B \end{cases}$

$n = 7$, $p_* = 3$



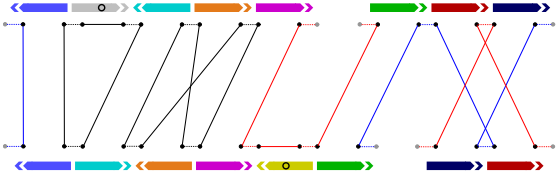
Capped relational graph

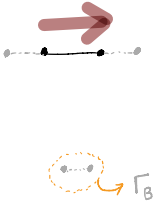
$n = \#$ common families

$p_* = \max \begin{cases} \#\text{ linear chromosomes in } A \\ \#\text{ linear chromosomes in } B \end{cases}$

$n = 7$, $p_* = 3$

6 cap vertices per genome





add $2p_*$ cap vertices to each genome

each cap is connected by an adjacency edge to the end of a path

if # linear chromosomes is different, connect each pair of isolated caps with an adjacency edge

Γ_A: represents an adjacency between caps in A

Γ_B: represents an adjacency between caps in B

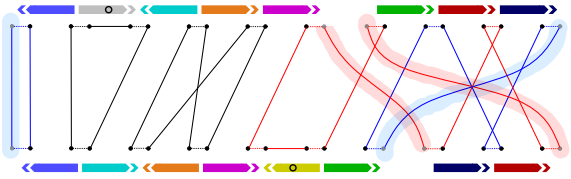
Capped relational graph

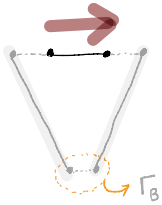
$n = \#$ common families

$p_* = \max \begin{cases} \#\text{ linear chromosomes in } A \\ \#\text{ linear chromosomes in } B \end{cases}$

$n = 7$, $p_* = 3$

6 cap vertices per genome





add $2p_*$ cap vertices to each genome

each cap is connected by an adjacency edge to the end of a path

link each cap from A to a cap from B: only cycles

if # linear chromosomes is different, connect each pair of isolated caps with an adjacency edge

Γ_A: represents an adjacency between caps in A

Γ_B: represents an adjacency between caps in B

DCJ-optimal capping

paths	linking cycle	Δn	$\Delta \lvert \mathcal{C} \rvert$	$\Delta(2\lvert \mathcal{P}_{AB} \rvert)$	Δ_{DCJ}
AB	(AB)	+0.5	+1	−0.5	0
AA + BB	(AA,BB)	+1	+1	0	0
AA	(AA,Γ_B)	+1	+1	0	0
BB	(BB,Γ_A)	+1	+1	0	0

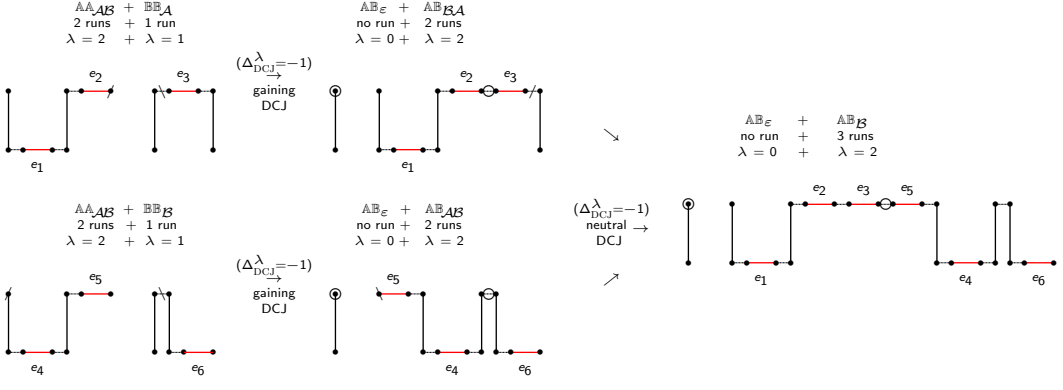
Closing an AA-path requires an adjacency Γ_B.

Closing a BB-path requires an adjacency Γ_A.

Capped relational graph of singular genomes - example

Deducting chain of path recombinations

- transforming $2 \times \mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{A}} + \mathbb{BB}_{\mathcal{B}}$
- into $3 \times \mathbb{AB}_{\varepsilon} + \mathbb{AB}_{\mathcal{B}}$
- with overall $\Delta^{\lambda}_{\mathrm{DCJ}} = -3$



DCJ-indel optimal capping

Indel-enclosing paths:	AB_{A}	AB_{B}	AB_{AB}	AB_{BA}
	AA_{A}	AA_{B}	AA_{AB} $=$	AA_{BA}
	BB_{A}	BB_{B}	BB_{AB} $=$	BB_{BA}

$P_{A} \simeq P_{ABA} \simeq P_{ABABA} \simeq \dots$

$P_{B} \simeq P_{BAB} \simeq P_{BABAB} \simeq \dots$

$P_{AB} \simeq P_{ABAB} \simeq P_{ABABAB} \simeq \dots$

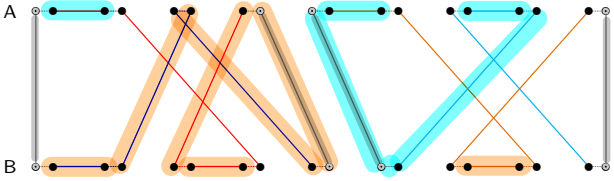
$P_{BA} \simeq P_{BABA} \simeq P_{BABABA} \simeq \dots$

DCJ-indel optimal capping

Indel-enclosing paths:	AB_A	AB_B	AB_{AB}	AB_{BA}
	AA_A	AA_B	AA_{AB}	$= AA_{BA}$
	BB_A	BB_B	BB_{AB}	$= BB_{BA}$

$P_A \simeq P_{ABA} \simeq P_{ABABA} \simeq \dots$
$P_B \simeq P_{BAB} \simeq P_{BABAB} \simeq \dots$
$P_{AB} \simeq P_{ABAB} \simeq P_{ABABAB} \simeq \dots$
$P_{BA} \simeq P_{BABA} \simeq P_{BABABA} \simeq \dots$

$n = 4$ and $p_* = 2$ paths: $2 \times AA_{AB}$, BB_A, BB_B



DCJ-optimal capping: 2 cycles (AA_{AB}, BB_B) and (BB_A, AA_{AB})

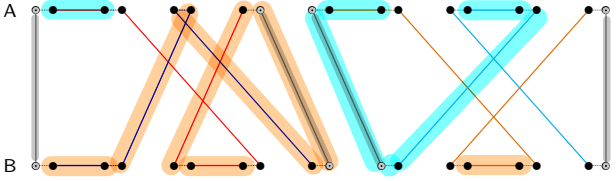
$$n + p_* - |\mathcal{C}| + \sum \lambda$$
$$= 4 + 2 - 2 + 4$$
$$= 8$$

DCJ-indel optimal capping

Indel-enclosing paths:	AB_A	AB_B	AB_{AB}	AB_{BA}
	AA_A	AA_B	AA_{AB}	$= AA_{BA}$
	BB_A	BB_B	BB_{AB}	$= BB_{BA}$

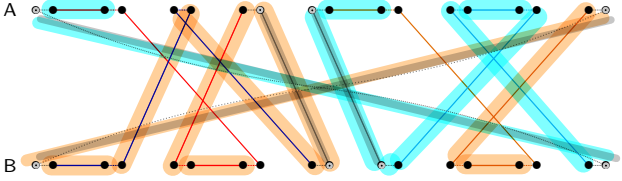
$P_A \simeq P_{ABA} \simeq P_{ABABA} \simeq \dots$
$P_B \simeq P_{BAB} \simeq P_{BABAB} \simeq \dots$
$P_{AB} \simeq P_{ABAB} \simeq P_{ABABAB} \simeq \dots$
$P_{BA} \simeq P_{BABA} \simeq P_{BABABA} \simeq \dots$

$n = 4$ and $p_* = 2$ paths: $2 \times AA_{AB}$, BB_A, BB_B



DCJ-optimal capping: 2 cycles
(AA_{AB}, BB_B) and (BB_A, AA_{AB})

$$n + p_* - |\mathcal{C}| + \sum \lambda$$
$$= 4 + 2 - 2 + 4$$
$$= 8$$

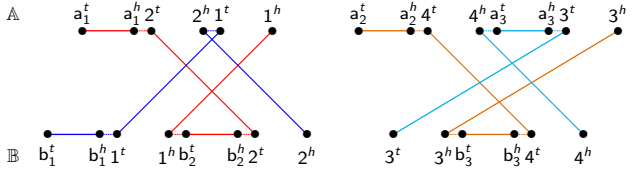


DCJ-indel optimal capping: one cycle
$(AA_{AB}, BB_B, AA_{BA}, BB_A)$

$$d_{\text{DCJ}}^{\text{ID}} = n + p_* - |\mathcal{C}| + \sum \lambda$$
$$= 4 + 2 - 1 + 2$$
$$= 7$$

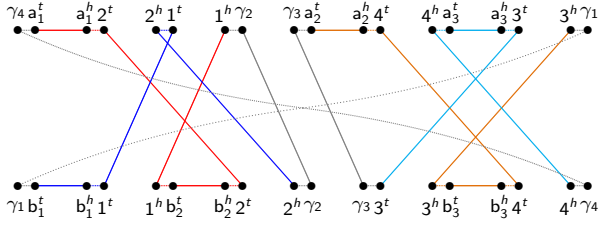
Capped relational graph of singular genomes - revisiting example

$\mathbb{A} = [\,a_1\ 2\ 1\,]\ [\,a_2\ 4\ a_3\ 3\,]$ and $\mathbb{B} = [\,b_1\ 1\ b_2\ 2\,]\ [\,3\ b_3\ 4\,]$; $p_* = 2$ and $a_* = 0$



Components: $2 \times \mathbb{AA}_{\mathcal{AB}}$, $\mathbb{BB}_{\mathcal{A}}$, $\mathbb{BB}_{\mathcal{B}}$

$$\begin{aligned} \mathrm{d}_{\mathrm{DCJ}}^{\mathrm{ID}} &= n - |\mathcal{C}| - \frac{|\mathcal{P}_{\mathbb{AB}}|}{2} + \sum \lambda(C) - \delta \\ &= 4 - 0 - 0 + 6 - 3 \\ &= 7 \end{aligned}$$



Linking cycle: $(\mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\mathcal{B}}, \mathbb{AA}_{\mathcal{BA}}, \mathbb{BB}_{\mathcal{A}})$

$$\begin{aligned} \mathrm{d}_{\mathrm{DCJ}}^{\mathrm{ID}} &= n + p_* - |\mathcal{C}| + \sum \lambda(C) \\ &= 4 + 2 - 1 + 2 \\ &= 7 \end{aligned}$$

The four sources of a chain of deducting recombinations are optimally linked into a single cycle.

Capped relational graph of singular genomes

The sources of each chain of deducting recombinations must be properly linked together into a single cycle.

- Unbalanced chains over-represented in genome $\mathbb{A}$ are marked with a $\cup$
  - $\mathbb{BB}_\varepsilon \prec \Gamma_\mathbb{B}$: a path $\mathbb{BB}_\varepsilon$ is preferred to close a $\cup$-unbalanced chain; if it does not exist, an adjacency $\Gamma_\mathbb{B}$ is used
- Unbalanced chains over-represented in genome $\mathbb{B}$ are marked with a $\cap$
  - $\mathbb{AA}_\varepsilon \prec \Gamma_\mathbb{A}$: a path $\mathbb{AA}_\varepsilon$ is preferred to close a $\cap$-unbalanced chain; if it does not exist, an adjacency $\Gamma_\mathbb{A}$ is used

In order to give the correct order of linking
- a path $\mathbb{AB}_{\mathcal{AB}}$ can be represented by $\mathbb{BA}_{\mathcal{BA}}$
- a path $\mathbb{AB}_{\mathcal{BA}}$ can be represented by $\mathbb{BA}_{\mathcal{AB}}$

DCJ-indel optimal capping

	id	sources	linking cycle		Δn	Δc	$\Delta(2\mathbb{AB})$	$\Delta\lambda$	$\Delta^{\lambda}_{\text{DCJ}}$
$\mathcal{P}$	WM	$\mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{AB}}$	$(\mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\mathcal{BA}})$		$+1$	$+1$	0	-2	-2
$\mathcal{Q}$	WW$\overline{\text{MM}}$	$2 \times \mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{A}} + \mathbb{BB}_{\mathcal{B}}$	$(\mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\mathcal{B}}, \mathbb{AA}_{\mathcal{BA}}, \mathbb{BB}_{\mathcal{A}})$		$+2$	$+1$	0	-4	-3
	MM$\overline{\text{WW}}$	$2 \times \mathbb{BB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{A}} + \mathbb{AA}_{\mathcal{B}}$	$(\mathbb{BB}_{\mathcal{AB}}, \mathbb{AA}_{\mathcal{B}}, \mathbb{BB}_{\mathcal{BA}}, \mathbb{AA}_{\mathcal{A}})$		$+2$	$+1$	0	-4	-3
$\mathcal{T}$	WZ$\overline{\text{M}}$	$\mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{A}} + \mathbb{AB}_{\mathcal{AB}}$	$(\mathbb{AB}_{\mathcal{AB}}, \mathbb{AA}_{\mathcal{BA}}, \mathbb{BB}_{\mathcal{A}})$		$+1.5$	$+1$	-0.5	-3	-2
	WW$\overline{\text{M}}$	$2 \times \mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{A}}$	$(\mathbb{AA}_{\mathcal{BA}}, \mathbb{BB}_{\mathcal{A}}, \mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\varepsilon} \prec \Gamma_{\mathbb{B}})$	$\cup$	$+2$	$+1$	0	-3	-2
	WNM	$\mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{B}} + \mathbb{AB}_{\mathcal{BA}}$	$(\mathbb{AB}_{\mathcal{BA}}, \mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\mathcal{B}})$		$+1.5$	$+1$	-0.5	-3	-2
	WW$\overline{\text{M}}$	$2 \times \mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{B}}$	$(\mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\mathcal{A}}, \mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\varepsilon} \prec \Gamma_{\mathbb{B}})$	$\cup$	$+2$	$+1$	0	-3	-2
	MN$\overline{\text{W}}$	$\mathbb{BB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{A}} + \mathbb{AB}_{\mathcal{BA}}$	$(\mathbb{AB}_{\mathcal{BA}}, \mathbb{AA}_{\mathcal{A}}, \mathbb{BB}_{\mathcal{AB}})$		$+1.5$	$+1$	-0.5	-3	-2
	MM$\overline{\text{W}}$	$2 \times \mathbb{BB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{A}}$	$(\mathbb{BB}_{\mathcal{BA}}, \mathbb{AA}_{\mathcal{A}}, \mathbb{BB}_{\mathcal{AB}}, \mathbb{AA}_{\varepsilon} \prec \Gamma_{\mathbb{A}})$	$\cap$	$+2$	$+1$	0	-3	-2
	MZ$\overline{\text{W}}$	$\mathbb{BB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{B}} + \mathbb{AB}_{\mathcal{AB}}$	$(\mathbb{AB}_{\mathcal{AB}}, \mathbb{AA}_{\mathcal{B}}, \mathbb{BB}_{\mathcal{BA}})$		$+1.5$	$+1$	-0.5	-3	-2
	MM$\overline{\text{W}}$	$2 \times \mathbb{BB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{B}}$	$(\mathbb{BB}_{\mathcal{AB}}, \mathbb{AA}_{\mathcal{B}}, \mathbb{BB}_{\mathcal{BA}}, \mathbb{AA}_{\varepsilon} \prec \Gamma_{\mathbb{A}})$	$\cap$	$+2$	$+1$	0	-3	-2

DCJ-indel optimal capping

	id	sources	linking cycle		Δn	Δc	$\Delta(2\mathbb{AB})$	$\Delta\lambda$	$\Delta^{\lambda}_{\mathrm{DCJ}}$
$\mathcal{S}$	ZN	$\mathbb{AB}_{\mathcal{AB}} + \mathbb{AB}_{\mathcal{BA}}$	$(\mathbb{AB}_{\mathcal{AB}}, \mathbb{AB}_{\mathcal{BA}})$		+1	+1	−1	−2	−1
	$\overline{\mathrm{WM}}$	$\mathbb{AA}_{\mathcal{A}} + \mathbb{BB}_{\mathcal{A}}$	$(\mathbb{AA}_{\mathcal{A}}, \mathbb{BB}_{\mathcal{A}})$		+1	+1	0	−1	−1
	$\underline{\mathrm{WM}}$	$\mathbb{AA}_{\mathcal{B}} + \mathbb{BB}_{\mathcal{B}}$	$(\mathbb{AA}_{\mathcal{B}}, \mathbb{BB}_{\mathcal{B}})$		+1	+1	0	−1	−1
	$\mathrm{W}\overline{\mathrm{M}}$	$\mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{A}}$	$(\mathbb{AA}_{\mathcal{BA}}, \mathbb{BB}_{\mathcal{A}})$		+1	+1	0	−1	−1
	$\mathrm{W}\underline{\mathrm{M}}$	$\mathbb{AA}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{B}}$	$(\mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\mathcal{B}})$		+1	+1	0	−1	−1
	WZ	$\mathbb{AA}_{\mathcal{AB}} + \mathbb{AB}_{\mathcal{AB}}$	$(\mathbb{AA}_{\mathcal{BA}}, \mathbb{BB}_{\varepsilon} \prec \Gamma_{\mathbb{B}}, \mathbb{AB}_{\mathcal{AB}})$	∪	+1.5	+1	−0.5	−2	−1
	WN	$\mathbb{AA}_{\mathcal{AB}} + \mathbb{AB}_{\mathcal{BA}}$	$(\mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\varepsilon} \prec \Gamma_{\mathbb{B}}, \mathbb{AB}_{\mathcal{BA}})$	∪	+1.5	+1	−0.5	−2	−1
	WW	$\mathbb{AA}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{AB}}$	$(\mathbb{AA}_{\mathcal{AB}}, \mathbb{BB}_{\varepsilon} \prec \Gamma_{\mathbb{B}}, \mathbb{AA}_{\mathcal{BA}}, \mathbb{BB}_{\varepsilon} \prec \Gamma_{\mathbb{B}})$	∪	+2	+1	0	−2	−1
	$\mathrm{M}\overline{\mathrm{W}}$	$\mathbb{BB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{A}}$	$(\mathbb{AA}_{\mathcal{A}}, \mathbb{BB}_{\mathcal{AB}})$		+1	+1	0	−1	−1
	$\mathrm{M}\underline{\mathrm{W}}$	$\mathbb{BB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{B}}$	$(\mathbb{AA}_{\mathcal{B}}, \mathbb{BB}_{\mathcal{BA}})$		+1	+1	0	−1	−1
	MZ	$\mathbb{BB}_{\mathcal{AB}} + \mathbb{AB}_{\mathcal{AB}}$	$(\mathbb{BB}_{\mathcal{BA}}, \mathbb{AB}_{\mathcal{AB}}, \mathbb{AA}_{\varepsilon} \prec \Gamma_{\mathbb{A}})$	∩	+1.5	+1	−0.5	−2	−1
	MN	$\mathbb{BB}_{\mathcal{AB}} + \mathbb{AB}_{\mathcal{BA}}$	$(\mathbb{BB}_{\mathcal{AB}}, \mathbb{AB}_{\mathcal{BA}}, \mathbb{AA}_{\varepsilon} \prec \Gamma_{\mathbb{A}})$	∩	+1.5	+1	−0.5	−2	−1
	MM	$\mathbb{BB}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{AB}}$	$(\mathbb{BB}_{\mathcal{AB}}, \mathbb{AA}_{\varepsilon} \prec \Gamma_{\mathbb{A}}, \mathbb{BB}_{\mathcal{BA}}, \mathbb{AA}_{\varepsilon} \prec \Gamma_{\mathbb{A}})$	∩	+2	+1	0	−2	−1
$\mathcal{M}$	ZZ$\underline{\mathrm{W}}\overline{\mathrm{M}}$	$2 \times \mathbb{AB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{B}} + \mathbb{BB}_{\mathcal{A}}$	$(\mathbb{AB}_{\mathcal{AB}}, \mathbb{AA}_{\mathcal{B}}, \mathbb{BA}_{\mathcal{BA}}, \mathbb{BB}_{\mathcal{A}})$		+2	+1	−1	−4	−2
	NN$\overline{\mathrm{W}}\underline{\mathrm{M}}$	$2 \times \mathbb{AB}_{\mathcal{BA}} + \mathbb{AA}_{\mathcal{A}} + \mathbb{BB}_{\mathcal{B}}$	$(\mathbb{AB}_{\mathcal{BA}}, \mathbb{AA}_{\mathcal{A}}, \mathbb{BA}_{\mathcal{AB}}, \mathbb{BB}_{\mathcal{B}})$		+2	+1	−1	−4	−2
$\mathcal{N}$	ZW$\overline{\mathrm{M}}$	$\mathbb{AB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{B}} + \mathbb{BB}_{\mathcal{A}}$	$(\mathbb{AB}_{\mathcal{AB}}, \mathbb{AA}_{\mathcal{B}}, \mathbb{BB}_{\mathcal{A}})$		+1.5	+1	−0.5	−2	−1
	ZZ$\underline{\mathrm{W}}$	$2 \times \mathbb{AB}_{\mathcal{AB}} + \mathbb{AA}_{\mathcal{B}}$	$(\mathbb{AB}_{\mathcal{AB}}, \mathbb{AA}_{\mathcal{B}}, \mathbb{BA}_{\mathcal{BA}}, \mathbb{BB}_{\varepsilon} \prec \Gamma_{\mathbb{B}})$	∪	+2	+1	−1	−3	−1
	ZZ$\overline{\mathrm{M}}$	$2 \times \mathbb{AB}_{\mathcal{AB}} + \mathbb{BB}_{\mathcal{A}}$	$(\mathbb{BA}_{\mathcal{BA}}, \mathbb{BB}_{\mathcal{A}}, \mathbb{AB}_{\mathcal{AB}}, \mathbb{AA}_{\varepsilon} \prec \Gamma_{\mathbb{A}})$	∩	+2	+1	−1	−3	−1
	N$\overline{\mathrm{W}}\underline{\mathrm{M}}$	$\mathbb{AB}_{\mathcal{BA}} + \mathbb{AA}_{\mathcal{A}} + \mathbb{BB}_{\mathcal{B}}$	$(\mathbb{AB}_{\mathcal{BA}}, \mathbb{AA}_{\mathcal{A}}, \mathbb{BB}_{\mathcal{B}})$		+1.5	+1	−0.5	−2	−1
	NN$\overline{\mathrm{W}}$	$2 \times \mathbb{AB}_{\mathcal{BA}} + \mathbb{AA}_{\mathcal{A}}$	$(\mathbb{AB}_{\mathcal{BA}}, \mathbb{AA}_{\mathcal{A}}, \mathbb{BA}_{\mathcal{AB}}, \mathbb{BB}_{\varepsilon} \prec \Gamma_{\mathbb{B}})$	∪	+2	+1	−1	−3	−1
	NN$\underline{\mathrm{M}}$	$2 \times \mathbb{AB}_{\mathcal{BA}} + \mathbb{BB}_{\mathcal{B}}$	$(\mathbb{BA}_{\mathcal{AB}}, \mathbb{BB}_{\mathcal{B}}, \mathbb{AB}_{\mathcal{BA}}, \mathbb{AA}_{\varepsilon} \prec \Gamma_{\mathbb{A}})$	∩	+2	+1	−1	−3	−1

DCJ-indel optimal capping

	remaining paths	linking cycle		Δn	Δc	$\Delta(2\mathbb{AB})$	$\Delta\lambda$	$\Delta^{\lambda}_{\text{DCJ}}$
1	$\mathbb{AB}_*$	$(\mathbb{AB}_*)$		$+0.5$	$+1$	-0.5	0	0
2	$\mathbb{AA}_* + \mathbb{BB}_*$	$(\mathbb{AA}_*, \mathbb{BB}_*)$		$+1$	$+1$	0	0	0
3	$\mathbb{AA}_*$	$(\mathbb{AA}_*, \Gamma_{\mathbb{B}})$	$\cup$	$+1$	$+1$	0	0	0
4	$\mathbb{BB}_*$	$(\mathbb{BB}_*, \Gamma_{\mathbb{A}})$	$\cap$	$+1$	$+1$	0	0	0

Any capping producing linking cycles following a top-down screening of the table above is optimal:

- $\Delta^{\lambda}_{\text{DCJ}} = \Delta n - \Delta c - \Delta(2\mathbb{AB}) + \Delta\lambda$ gives the DCJ-indel-effect produced by each type of linking cycle.
- All given linking cycles have $\Delta^{\lambda}_{\text{DCJ}}$ equivalent to the respective chain of deducting recombinations, therefore they achieve the optimal DCJ-indel distance.

P1: After identifying chains of recombinations $\begin{cases} \text{either there are no unbalanced chains} \\ \text{or there are only } \cup\text{-unbalanced chains (over-repr. in } \mathbb{A}) \\ \text{or there are only } \cap\text{-unbalanced chains (over-repr. in } \mathbb{B}) \end{cases}$

P2: When an unbalanced chain is being linked

$\begin{cases} \text{if there is a remaining indel-free } \mathbb{AA}_{\varepsilon}/\mathbb{BB}_{\varepsilon} \text{ (of the under-repr. genome), it is used to link the chain} \\ \text{otherwise there is no remaining } \mathbb{AA}_*/\mathbb{BB}_* \text{ (of the under-repr. genome) and an adjacency } \Gamma_{\mathbb{A}/\mathbb{B}} \text{ links the chain} \end{cases}$

Any optimal capping that links all possible chains of deducting recombinations as described above and, for the remaining paths, maximizes the number of linking cycles of type **2** minimizes the number of caps:

$\begin{cases} \text{The number of caps to be added is exactly } p_* = \max\{\kappa_{\mathbb{A}}, \kappa_{\mathbb{B}}\}\,. \\ \text{The number of adjacencies between caps is exactly } a_* = |\kappa_{\mathbb{A}} - \kappa_{\mathbb{B}}|. \end{cases}$

Quiz 2

1 Which of the following statements about the capped relational graph are true?

A In an optimal capping, the distance computed based on the capped relational diagram must be equivalent to the distance computed based on the original relational diagram.

B Let $RG(\mathbb{A}, \mathbb{B})$ be a relational graph of **canonical** genomes.
An optimal capping of $RG(\mathbb{A}, \mathbb{B})$ that maximizes the number of cycles linking a pair $\mathbb{A}\mathbb{A} + \mathbb{B}\mathbb{B}$ has a minimum number of caps $(= \max\{\kappa_{\mathbb{A}}, \kappa_{\mathbb{B}}\})$.

C Let $\max\{\kappa_{\mathbb{A}_s}, \kappa_{\mathbb{B}_s}\} = \max\{\kappa_{\mathbb{A}_c}, \kappa_{\mathbb{B}_c}\}$.
An optimal capping of the relational graph of **singular** genomes $\mathbb{A}_s$ and $\mathbb{B}_s$ requires more caps than an optimal capping of the relational graph of **canonical** genomes $\mathbb{A}_c$ and $\mathbb{B}_c$.

D The indel-potential can be equivalently computed based on the number of runs or based on the number of transitions.

References

Double Cut and Join with Insertions and Deletions

(Marília D.V. Braga, Eyla Willing and Jens Stoye)

JCB, Vol. 18, No. 9 (2011)

Sorting Linear Genomes with Rearrangements and Indels

(Marília D. V. Braga and Jens Stoye)

TCBB, vol 12, issue 3, pp. 500-506 (2015)

On the weight of indels in genomic distances

(Marília D. V. Braga, Raphael Machado, Leonardo C. Ribeiro and Jens Stoye)

BMC Bioinformatics, vol. 12, S13 (2011)

Computing the Rearrangement Distance of Natural Genomes

(Leonard Bohnenkämper, Marília D. V. Braga, Daniel Doerr and Jens Stoye)

JCB, Vol. 28, No. 4 (2021)