## Topics of today:

1. Canonical DCJ distance and sorting

2. Relational graph

3. Restricted canonical DCJ sorting

# Canonical DCJ

Given two canonical genomes $\mathbb{A}$ and $\mathbb{B}$,...

**Canonical DCJ Distance Problem:** Compute the minimum number of DCJ operations required to transform $\mathbb{A}$ into $\mathbb{B}$.

Denote by $d_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B})$ the DCJ distance of $\mathbb{A}$ and $\mathbb{B}$.

**Canonical DCJ Sorting Problem:** Find a sequence of $d_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B})$ DCJ operations that transform $\mathbb{A}$ into $\mathbb{B}$.

Let $\xi(\mathbb{G}) = \{g^t : g \in \mathcal{G}(\mathbb{G})\} \cup \{g^h : g \in \mathcal{G}(\mathbb{G})\}$ be the set of extremities of all genes in genome $\mathbb{G}$.

Ex: $\mathbb{G} = [2\,\bar{3}\,1]$, $\mathcal{G}(\mathbb{G}) = \{1, 2, 3\}$ and $\xi(\mathbb{G}) = \{1^t, 1^h, 2^t, 2^h, 3^t, 3^h\}$.

Note that, if genomes $\mathbb{A}$ and $\mathbb{B}$ are canonical, then $\xi(\mathbb{A}) = \xi(\mathbb{B})$.

# Relational graph of canonical genomes

Given two canonical genomes $\mathbb{A}$ and $\mathbb{B}$, their **relational graph** $RG(\mathbb{A}, \mathbb{B}) = (V, E)$ is described as follows:

1. $V = V(\xi(\mathbb{A})) \cup V(\xi(\mathbb{B}))$ :    there is a vertex for each extremity of each gene in $\mathbb{A}$
   and a vertex for each extremity of each gene in $\mathbb{B}$

   Each vertex $v$ has a label $\ell(v)$, that corresponds to the extremity it represents.

   For a given gene $g$, let $\begin{cases} \text{vertices } u \text{ and } v \text{ represent } g^t \text{ and } g^h \text{ in genome } \mathbb{A} \text{ and} \\ \text{vertices } u' \text{ and } v' \text{ represent } g^t \text{ and } g^h \text{ in genome } \mathbb{B} \end{cases}$

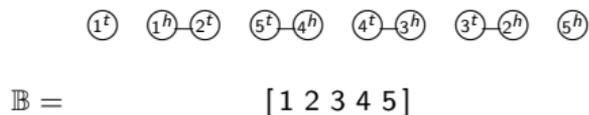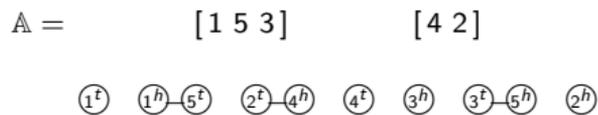   Then: $\ell(u) = \ell(u') = g^t$ and $\ell(v) = \ell(v') = g^h$

2. $E = E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B}) \cup E_\xi$, where:

   - **Adjacency edges:** $\begin{cases} E_\alpha(\mathbb{A}) = \{uv : u, v \in V(\xi(\mathbb{A})) \text{ and } \ell(u)\ell(v) \in \alpha(\mathbb{A})\} \\ E_\alpha(\mathbb{B}) = \{uv : u, v \in V(\xi(\mathbb{B})) \text{ and } \ell(u)\ell(v) \in \alpha(\mathbb{B})\} \end{cases}$

   - **Extremity edges:** $E_\xi = \{uv : u \in V(\xi(\mathbb{A})) \text{ and } v \in V(\xi(\mathbb{B})) \text{ and } \ell(u) = \ell(v)\}$

Note that:

- Let $n = |\mathcal{G}_\star|$. The number of edges in $E_\xi$ is $2n$ (two edges per element of $\mathcal{G}_\star$).

# Relational graph of canonical genomes

$\mathbb{A} =$       [1 5 3]      [4 2]

$\text{①}^t$   $\text{①}^h\!\!-\!\!\text{⑤}^t$   $\text{②}^t\!\!-\!\!\text{④}^h$   $\text{④}^t$   $\text{③}^h$   $\text{③}^t\!\!-\!\!\text{⑤}^h$   $\text{②}^h$

$\text{①}^t$   $\text{①}^h\!\!-\!\!\text{②}^t$   $\text{⑤}^t\!\!-\!\!\text{④}^h$   $\text{④}^t\!\!-\!\!\text{③}^h$   $\text{③}^t\!\!-\!\!\text{②}^h$   $\text{⑤}^h$

$\mathbb{B} =$       [1 2 3 4 5]

$$n = |\mathcal{G}_\star| = 5 \ , \quad \kappa(\mathbb{A}) = 2 \quad \text{and} \quad \kappa(\mathbb{B}) = 1$$

---

Every vertex has degree one or two:
$RG(\mathbb{A}, \mathbb{B})$ is a collection of paths and cycles
(alternating edes in $E_\xi$ and in $E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$)

cycle with $k$ edges in $E_\xi$: $k$-cycle or $c_k$
path with $k$ edges n $E_\xi$: $k$-path or $p_k$

$$\begin{cases} \mathcal{C} = \{c_k\} \ : \ \text{set of cycles ($k$ is even)} \\ \mathcal{P}_{\mathbb{AA}} = \{p_k : \text{starts and ends in } \mathbb{A}\} : \\ \qquad\qquad \text{set of } \mathbb{AA}\text{-paths ($k$ is even)} \\ \mathcal{P}_{\mathbb{BB}} = \{p_k : \text{starts and ends in } \mathbb{B}\} : \\ \qquad\qquad \text{set of } \mathbb{BB}\text{-paths ($k$ is even)} \\ \mathcal{P}_{\mathbb{AB}} = \{p_k : \text{starts in } \mathbb{A} \text{ and ends in } \mathbb{B}\} : \\ \qquad\qquad \text{set of } \mathbb{AB}\text{-paths ($k$ is odd)} \end{cases}$$

$|\mathcal{P}_{\mathbb{AB}}|$ is even ($E_\xi$ has $2n$ edges)
$|\mathcal{P}_{\mathbb{AA}}| + |\mathcal{P}_{\mathbb{BB}}| + |\mathcal{P}_{\mathbb{AB}}| = \kappa(\mathbb{A}) + \kappa(\mathbb{B})$
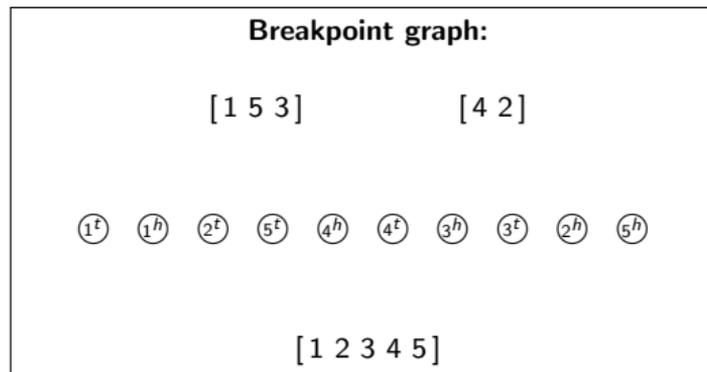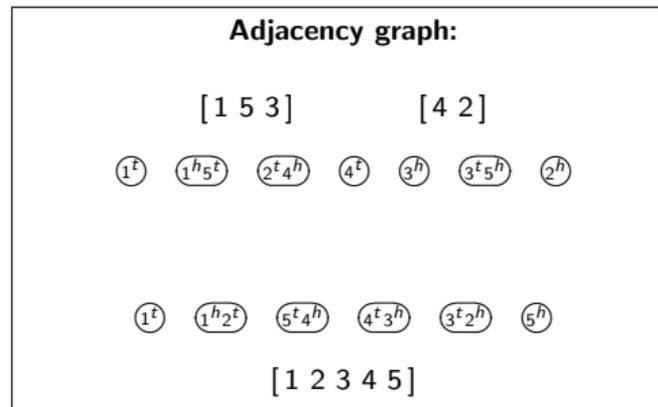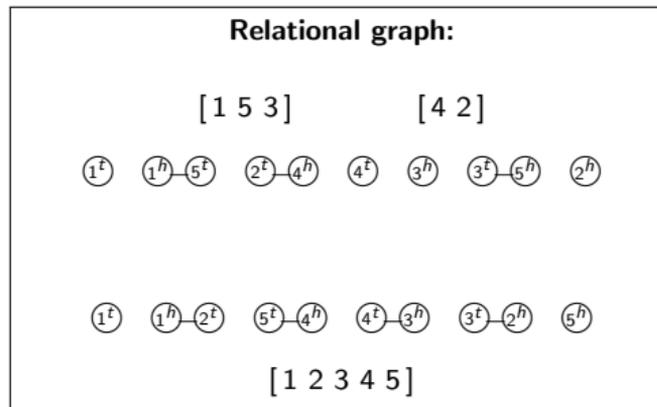
---

If $\mathbb{A} = \mathbb{B}$,
$RG(\mathbb{A}, \mathbb{B})$ has only 2-cycles and 1-paths:
$$2n = 2|\mathcal{C}| + |\mathcal{P}_{\mathbb{AB}}| \ \Rightarrow \ n = |\mathcal{C}| + \frac{|\mathcal{P}_{\mathbb{AB}}|}{2}$$

---

Otherwise, if $\mathbb{A} \neq \mathbb{B}$:
$$n > |\mathcal{C}| + \frac{|\mathcal{P}_{\mathbb{AB}}|}{2}$$

# Relational graph ≅ Adjacency graph ≅ Breakpoint graph

**Relational graph:**

[ 1 5 3]          [ 4 2 ]

$1^t$   $1^h$–$5^t$   $2^t$–$4^h$   $4^t$   $3^h$   $3^t$–$5^h$   $2^h$

$1^t$   $1^h$–$2^t$   $5^t$–$4^h$   $4^t$–$3^h$   $3^t$–$2^h$   $5^h$

[ 1 2 3 4 5 ]

**Adjacency graph:**

[ 1 5 3]          [ 4 2 ]

$1^t$   $1^h 5^t$   $2^t 4^h$   $4^t$   $3^h$   $3^t 5^h$   $2^h$

$1^t$   $1^h 2^t$   $5^t 4^h$   $4^t 3^h$   $3^t 2^h$   $5^h$

[ 1 2 3 4 5 ]

**Breakpoint graph:**

[ 1 5 3]               [ 4 2 ]

$1^t$   $1^h$   $2^t$   $5^t$   $4^h$   $4^t$   $3^h$   $3^t$   $2^h$   $5^h$

[ 1 2 3 4 5 ]

# Types of DCJ operation

Let a DCJ operation transform a genome $\mathbb{A}$ into another genome $\mathbb{A}'$:

$\left.\begin{array}{l} m_1 : \# \text{ of components in } RG(\mathbb{A}, \mathbb{B}) \\ m_2 : \# \text{ of components in } RG(\mathbb{A}', \mathbb{B}) \end{array}\right\} \quad 0 \le |m_2 - m_1| \le 1$

Goal: increase the number of cycles ($|\mathcal{C}|$) and/or the number of $\mathbb{A}\mathbb{B}$-paths ($|\mathcal{P}_{\mathbb{A}\mathbb{B}}|$) in $RG$

# Types of DCJ operation

Goal: increase the number of cycles ($|\mathcal{C}|$) and/or the number of $\mathbb{AB}$-paths ($|\mathcal{P}_{\mathbb{AB}}|$) in $RG$

# Types of DCJ operation

Goal: increase the number of cycles ($|\mathcal{C}|$) and/or the number of $\mathbb{AB}$-paths ($|\mathcal{P}_{\mathbb{AB}}|$) in $RG$
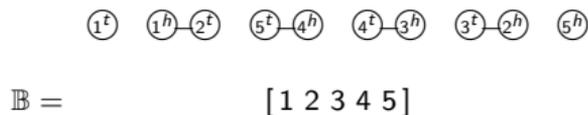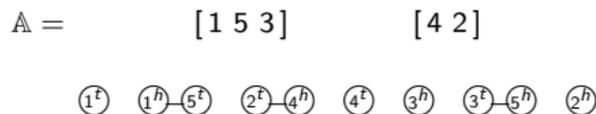
# Canonical DCJ Distance & Sorting

Recall that, if $\mathbb{A} = \mathbb{B}$, we have $n = |\mathcal{C}| + \frac{|\mathcal{P}_{\mathbb{A}\mathbb{B}}|}{2}$, otherwise $n > |\mathcal{C}| + \frac{|\mathcal{P}_{\mathbb{A}\mathbb{B}}|}{2}$

A DCJ operation $\rho$ is called **optimal** if $\left\{ \begin{array}{l} \rho \text{ increases the number of cycles by one, or} \\ \\ \rho \text{ increases the number of } \mathbb{A}\mathbb{B}\text{-paths by two.} \end{array} \right.$

Given two canonical genomes $\mathbb{A}$ and $\mathbb{B}$, it is possible to find an optimal DCJ operation at each sorting step. Therefore:

$$d_{\mathrm{DCJ}}(\mathbb{A}, \mathbb{B}) = n - |\mathcal{C}| - \frac{|\mathcal{P}_{\mathbb{A}\mathbb{B}}|}{2}$$

# Computing the canonical DCJ Distance in linear time

**A.** Telomeres and adjacencies of genome $\mathbb{A}$

| Pos | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-----|---|---|---|---|---|---|---|
| 1st | $1^t$ | $1^h$ | $5^h$ | $3^h$ | $4^t$ | $h^h$ | $2^h$ |
| 2nd | - | $5^t$ | $3^t$ | - | - | $2^t$ | - |

Positions of gene extremities in Tab. **A**

| gene | 1 | 2 | 3 | 4 | 5 |
|------|---|---|---|---|---|
| head | 2 | 7 | 4 | 6 | 3 |
| tail | 1 | 6 | 3 | 5 | 2 |

$\mathbb{A} = \qquad [1\ 5\ 3] \qquad [4\ 2]$

$①^t \quad ①^h-⑤^t \quad ②^t-④^h \quad ④^t \quad ③^h \quad ③^t-⑤^h \quad ②^h$

$①^t \quad ①^h-②^t \quad ⑤^t-④^h \quad ④^t-③^h \quad ③^t-②^h \quad ⑤^h$

$\mathbb{B} = \qquad [1\ 2\ 3\ 4\ 5]$

**B.** Telomeres and adjacencies of genome $\mathbb{B}$

| Pos | 1 | 2 | 3 | 4 | 5 | 6 |
|-----|---|---|---|---|---|---|
| 1st | $1^t$ | $1^h$ | $2^h$ | $3^h$ | $4^h$ | $5^h$ |
| 2nd | - | $2^t$ | $3^t$ | $4^t$ | $5^t$ | - |

Positions of gene extremities in Tab. **B**

| gene | 1 | 2 | 3 | 4 | 5 |
|------|---|---|---|---|---|
| head | 2 | 3 | 4 | 5 | 6 |
| tail | 1 | 2 | 3 | 4 | 5 |

# Canonical DCJ Sorting

---

**Algorithm 2** (Greedy sorting by DCJ)

---

 1: **for each** adjacency $\{p, q\}$ in genome $B$ **do**
 2:  let $u$ be the element of genome $A$ that contains $p$
 3:  let $v$ be the element of genome $A$ that contains $q$
 4:  **if** $u \neq v$ **then**
 5:   replace $u$ and $v$ in $A$ by $\{p, q\}$ and $(u \setminus \{p\}) \cup (v \setminus \{q\})$
 6:  **end if**
 7: **end for**
 8: **for each** telomere $\{p\}$ in genome $B$ **do**
 9:  let $u$ be the element of genome $A$ that contains $p$
10:  **if** $u$ is an adjacency **then**
11:   replace $u$ in $A$ by $\{p\}$ and $(u \setminus \{p\})$
12:  **end if**
13: **end for**

---

# Canonical DCJ Sorting

$\mathbb{A} =$     [1 5 3]     [4 2]

**A.** Telomeres and adjacencies of genome $\mathbb{A}$

| Pos | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-----|-----|-----|-----|-----|-----|-----|-----|
| 1st | $1^t$ | $1^h$ | $5^h$ | $3^h$ | $4^t$ | $h^h$ | $2^h$ |
| 2nd | - | $5^t$ | $3^t$ | - | - | $2^t$ | - |

Positions of gene extremities in Tab. **A**

| gene | 1 | 2 | 3 | 4 | 5 |
|------|---|---|---|---|---|
| head | 2 | 7 | 4 | 6 | 3 |
| tail | 1 | 6 | 3 | 5 | 2 |

**B.** Telomeres and adjacencies of genome $\mathbb{B}$

| Pos | 1 | 2 | 3 | 4 | 5 | 6 |
|-----|-----|-----|-----|-----|-----|-----|
| 1st | $1^t$ | $1^h$ | $2^h$ | $3^h$ | $4^h$ | $5^h$ |
| 2nd | - | $2^t$ | $3^t$ | $4^t$ | $5^t$ | - |

Positions of gene extremities in Tab. **B**

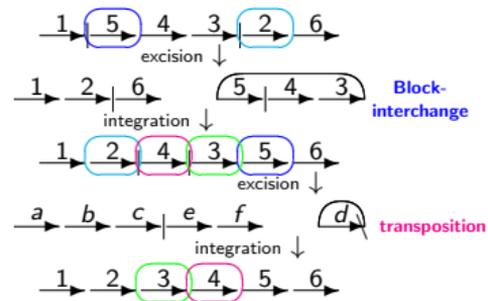| gene | 1 | 2 | 3 | 4 | 5 |
|------|---|---|---|---|---|
| head | 2 | 3 | 4 | 5 | 6 |
| tail | 1 | 2 | 3 | 4 | 5 |

$\mathbb{B} =$     [1 2 3 4 5]

# DCJ model - circular excision/integration



**Canonical DCJ model**
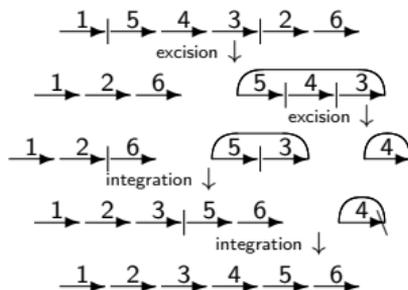
Many circular chromosomes can coexist
in the intermediate genomes.
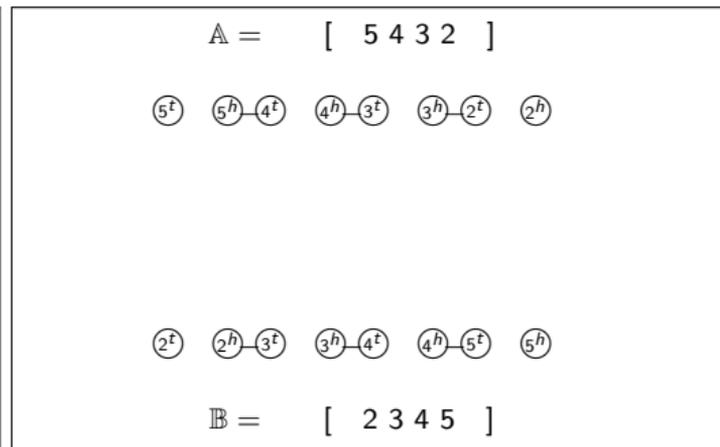
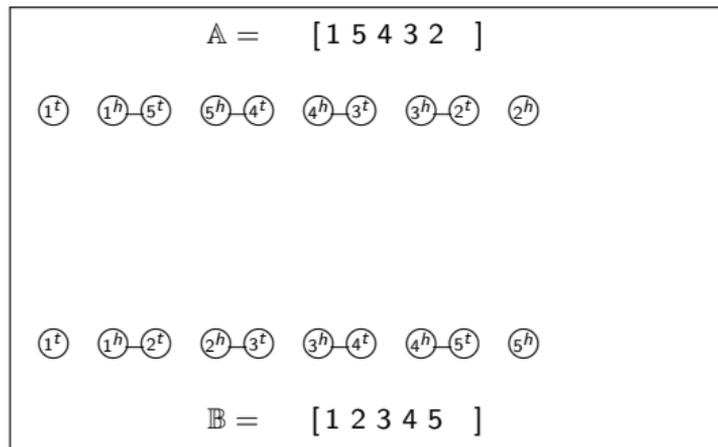**Restricted canonical DCJ model**

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 canonical DCJ sorting

Canonical genomes $\mathbb{A}$ and $\mathbb{B}$ are linear.

Transform $\mathbb{A}$ and $\mathbb{B}$ into **co-tailed** genomes $\mathbb{A}'$ and $\mathbb{B}'$:



Assume that the genes in chromosomes of $\mathbb{B}'$ are consecutive numbers
(otherwise renumber the genes of $\mathbb{A}'$ and of $\mathbb{B}'$)

# Restricted canonical DCJ sorting

Canonical genomes $\mathbb{A}'$ and $\mathbb{B}'$ are linear and co-tailed.

$\mathbb{A} = [\,1\ 5\ 4\ 3\ 2\ 6\,]$

Sort $\mathbb{A}'$ into $\mathbb{B}'$ from left to right:

Once we have transformed the beginning of a chromosome in $\mathbb{A}'$ to $k\ k{+}i\ \ldots \ell$, we extend it by moving $\ell{+}1$ next to $\ell$. Cases:

1. $\ell{+}1$ is on a different chromosome: translocation

Otherwise the situation is $\ell \ldots \ell{+}1 \ldots$

2. $\ell$ and $\ell{+}1$ have distinct orientations: inversion

Otherwise, find the highest gene $m$ between $\ell$ and $\ell{+}1$ and find $m{+}1$

3. $m{+}1$ is on a different chromosome: translocation to move it next to $m$; this operation also moves $\ell{+}1$ to another chromosome, with another translocation we put it next to $\ell$

Otherwise the situation is $\ell \ldots m \ldots \ell{+}1 \ldots m{+}1$

4. $m$ and $m{+}1$ have distinct orientations: inversion to move $m{+}1$ next to $m$; this also changes the orientation of $\ell{+}1$, with another inversion we put it next to $\ell$

5. $m$ and $m{+}1$ have the same orientation: interchange blocks

$\ell \, \langle \ldots m \rangle \ldots \langle \ell{+}1 \ldots \rangle \, m{+}1 \rightsquigarrow \ell \, \ell{+}1 \ldots m \, m{+}1$

(if both $m$ and $m{+}1$ have positive orientation)

or

$\ell \, \langle \ldots \rangle \, \overline{m} \ldots \langle \ell{+}1 \ldots \overline{m{+}1} \rangle \rightsquigarrow \ell \, \ell{+}1 \ldots \overline{m{+}1}\ \overline{m}$

(if $m$ and $m{+}1$ have both reverse orientation)

With two operations we put $\ell{+}1$ next to $\ell$ and $m$ next to $m{+}1$.

$\mathbb{B} = [\,1\ 2\ 3\ 4\ 5\ 6\,]$

# References

A Unifying View of Genome Rearrangements

(Anne Bergeron, Julia Mixtacki and Jens Stoye)

LNCS, volume 4175, pages 163-173 (2006)


Restricted DCJ Model: Rearrangement Problems with Chromosome Reincorporation

(Jakub Kováč, Robert Warren, Marília D. V. Braga, and Jens Stoye)

JCB Volume 18, Number 9, 2011