Topics of today:

1. DCJ double distance
2. DCJ halving
3. DCJ median
4. Inversion distance (unichromosomal genomes)

DCJ double distance

The DCJ distance of **balanced genomes** $\mathbb{A}^f$ and $\mathbb{B}^f$ is:

$$d_{DCJ}(\mathbb{A}^f, \mathbb{B}^f) = \min_{f_m} d_{DCJ}(\mathbb{A}_{\triangleright}^{f_m}, \mathbb{B}_{\triangleright}^{f_m})$$

where f_m is any function that produces a maximal matching of the families defined by f

DCJ double distance:

$$d_{DCJ}^2(\mathbb{S}_{\triangleright}^f, \mathbb{D}_{\diamond}^f) = \min_{\mathbb{P}_{\bowtie}^f \in 2\cdot\mathbb{S}_{\triangleright}^f} d_{DCJ}(\mathbb{P}_{\bowtie}^f, \mathbb{D}_{\diamond}^f)$$

DCJ halving

DCJ Halving Distance Problem:

Compute the minimum DCJ double distance for a (rearranged) duplicated genome $\mathbb{D}^f$:

$$\mathsf{h}_{\mathrm{DCJ}}(\mathbb{D}^f) = \min_{\text{singular } \mathbb{H}^f} \mathsf{d}^2_{\mathrm{DCJ}}(\mathbb{H}^f, \mathbb{D}^f)$$

DCJ Halving Problem:

Find a singular genome $\widehat{\mathbb{H}^f}$ and a perfectly duplicated genome $\widehat{\mathbb{P}^f} \in 2{\cdot}\widehat{\mathbb{H}^f}$ such that

$$\mathsf{h}_{\mathrm{DCJ}}(\mathbb{D}^f) = \mathsf{d}^2_{\mathrm{DCJ}}(\widehat{\mathbb{H}^f}, \mathbb{D}^f) = \mathsf{d}_{\mathrm{DCJ}}(\widehat{\mathbb{P}^f}, \mathbb{D}^f)$$

DCJ Halving Sorting Problem:

Give a sequence of $\mathsf{h}_{\mathrm{DCJ}}(\mathbb{D}^f)$ DCJ operations that transform $\mathbb{D}^f$ into $\widehat{\mathbb{P}^f}$

Can be solved at once, also giving the matching between $\mathbb{D}^f$ and $\widehat{\mathbb{P}^f}$

Natural graph of a duplicated genome

Natural graph $NG(\mathbb{D}^f) = (V, E)$ of a duplicated genome $\mathbb{D}^f$:

First assign arbitrarily indices $\mathtt{a}$ and $\mathtt{b}$ to the two genes of each family in $\mathbb{D}^f$, obtaining $\mathbb{D}'^f$

1. $V = \Gamma(\mathbb{D}'^f) \cup \Theta(\mathbb{D}'^f)$ (each adjacency or telomere of $\mathbb{D}'^f$ is a vertex of $NG(\mathbb{D}^f)$)
2. For each family $\mathtt{X} \in \mathcal{F}(\mathbb{D}^f)$, each pair of paralogous extremities is connected by an edge in $NG(\mathbb{D}^f)$, i.e.:
   - there is an edge connecting the vertex u that contain $\mathtt{X}_\mathtt{a}^h$ and the vertex v that contain $\mathtt{X}_\mathtt{b}^h$
   - there is an edge connecting the vertex u' that contain $\mathtt{X}_\mathtt{a}^t$ and the vertex v that contain $\mathtt{X}_\mathtt{b}^t$

Note that:

- There can be adjacencies/vertices of type $\mathtt{X}_\mathtt{a}^h\mathtt{X}_\mathtt{b}^h$ and/or $\mathtt{X}_\mathtt{a}^t\mathtt{X}_\mathtt{b}^t$ ($NG(\mathbb{D}^f)$ can contain 1-cycles)
- Let $n = |\mathcal{F}(\mathbb{D}^f)| = \frac{|\mathcal{G}(\mathbb{D}^f)|}{2}$. The number of edges in $NG(\mathbb{D}^f) = 2n$ (two edges per element of $\mathcal{F}(\mathbb{D}^f)$).

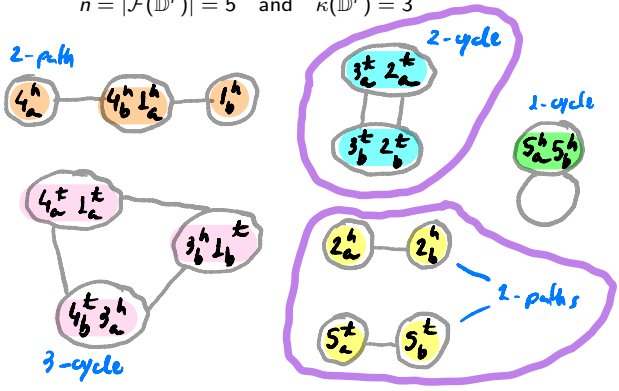
Natural graph of a duplicated genome

Ex: $\mathbb{D}'^f = [\bar{4}_a\ 1_a\ \bar{4}_b\ \bar{3}_a\ 2_a] \quad [\bar{2}_b\ 3_b\ 1_b] \quad [5_a\ \bar{5}_b]$

1-paths and 2-cycles are sorted

$\Gamma(\mathbb{D}) \cup \Theta(\mathbb{D}) = \{ 4_a^h, 4_a^t 1_a^t, 1_a^h 4_b^h, 4_b^t 3_a^h, 3_a^t 2_a^t, 2_a^h, 2_b^h, 2_b^t 3_b^t, 3_b^h 1_b^t, 1_b^h, 5_a^t, 5_a^h 5_b^h, 5_b^t \}$

$n = |\mathcal{F}(\mathbb{D}^f)| = 5$ and $\kappa(\mathbb{D}^f) = 3$



Every vertex has degree one or two:
$NG(\mathbb{D}^f)$ is a collection of paths and cycles

cycle with k edges: k-cycle or c_k
path with k edges: k-path or p_k

$$\begin{cases} \mathcal{C}_e = \{c_k : k \text{ is even}\} & : \text{ set of even cycles} \\ \mathcal{P}_e = \{p_k : k \text{ is even}\} & : \text{ set of even paths} \\ \mathcal{C}_o = \{c_k : k \text{ is odd}\} & : \text{ set of odd cycles} \\ \mathcal{P}_o = \{p_k : k \text{ is odd}\} & : \text{ set of odd paths} \end{cases}$$

$|\mathcal{C}_o| + |\mathcal{P}_o|$ is even (NG has $2n$ edges)
$|\mathcal{P}_e| + |\mathcal{P}_o| = \kappa(\mathbb{D}^f)$

For a perfectly duplicated genome $\mathbb{P}^f$,
$NG(\mathbb{P}^f)$ has only 2-cycles and 1-paths:

$$2n = 2|\mathcal{C}_e| + |\mathcal{P}_o| \quad \Rightarrow \quad n = |\mathcal{C}_e| + \frac{|\mathcal{P}_o|}{2}$$

Otherwise, if a duplicated genome $\mathbb{D}^f$
is not perfectly duplicated:

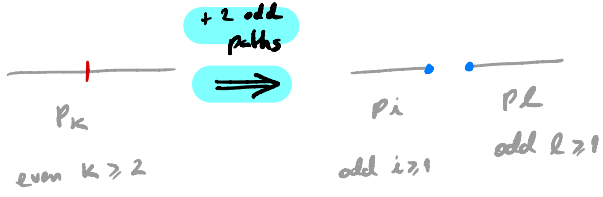
$$n > |\mathcal{C}_e| + \left\lfloor \frac{|\mathcal{P}_o|}{2} \right\rfloor$$

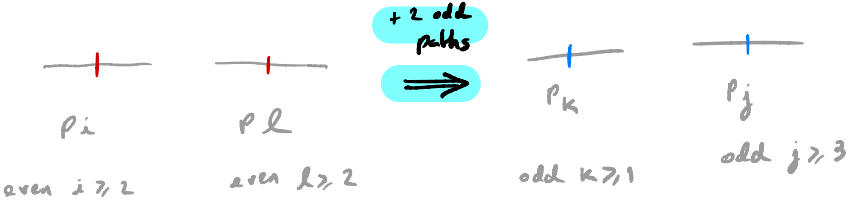
Types of DCJ operation

Goal: increase the number of even cycles ($|\mathcal{C}_e|$) and/or the number of odd paths ($|\mathcal{P}_o|$) in NG

Types of DCJ operation

Goal: increase the number of even cycles ($|\mathcal{C}_e|$) and/or odd paths ($|\mathcal{P}_o|$) in NG





DCJ Halving: Sorting & Distance

Recall that, if the genome is perfectly duplicated, we have $n = |\mathcal{C}_e| + \frac{|\mathcal{P}_o|}{2}$, otherwise $n > |\mathcal{C}_e| + \left\lfloor \frac{|\mathcal{P}_o|}{2} \right\rfloor$

A DCJ operation ρ is called **optimal** if

- ρ increases the number of even cycles by one, or
- ρ increases the number of odd paths by two, or
- the number of odd paths is odd and
  - ρ increases the number of odd paths by one

  (can occur at most once)

Given a duplicated genome $\mathbb{D}^f$, it is possible to find an optimal DCJ operation at each sorting step. Therefore:

$$\mathrm{h}_{\mathrm{DCJ}}(\mathbb{D}^f) = n - |\mathcal{C}_e| - \left\lfloor \frac{|\mathcal{P}_o|}{2} \right\rfloor$$

DCJ halving: obtaining an optimal perfectly duplicated genome

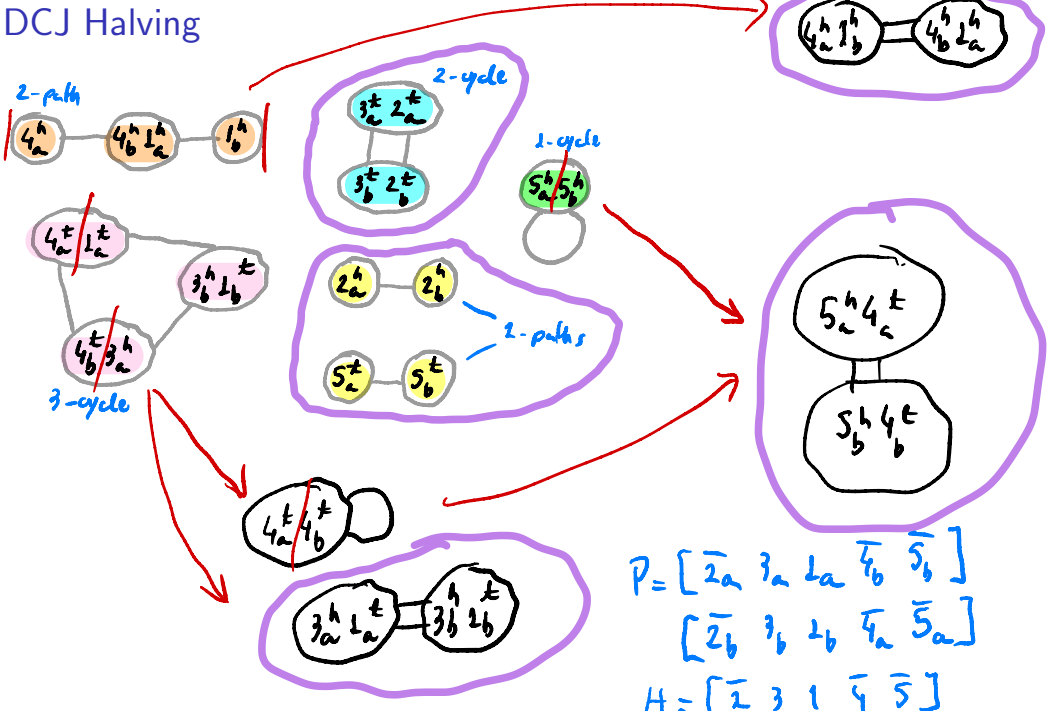
Given a duplicated genome $\mathbb{D}^f$,

with natural graph $NG(\mathbb{D}^f)$,

and DCJ halving distance $h = \mathrm{h}_{\mathrm{DCJ}}(\mathbb{D}^f) = n - |\mathcal{C}_e| - \left\lfloor \frac{|\mathcal{P}_o|}{2} \right\rfloor$:

1. $NG_0 \leftarrow NG(\mathbb{D}^f)$
2. For $i = 1$ to h :
   - Find and apply one optimal DCJ operation, transforming NG_{i-1} into NG_i.
3. NG_h is a simple collection of 2-cycles and 1-paths:
   reconstruct the perfectly duplicated genome $\mathbb{P}^f \in 2\cdot\mathbb{H}$ from NG_h.

DCJ Halving



$$P = [\,\bar{2}_a\ 3_a\ 1_a\ \bar{4}_b\ \bar{5}_b\,]$$
$$[\,\bar{2}_b\ 3_b\ 1_b\ \bar{4}_a\ \bar{5}_a\,]$$
$$H = [\,\bar{2}\ 3\ 1\ \bar{4}\ \bar{5}\,]$$

Quiz 1

1 Which of the following statements about the Natural Graph are true?

A Merging two odd cycles is always optimal.

B Breaking an odd cycle into an odd path cannot be optimal. optimal if |P_o| is odd

C Breaking an even path into two odd paths is always optimal.

D Breaking an even cycle into two cycles is always optimal. only if breaks into 2 even cycles

E Recombining two even paths into two odd paths is always optimal.

Solving the DCJ double distance

Let $\mathbb{S}^f$ be a singular and $\mathbb{D}^f$ be a duplicated genome.

We want to compute the double distance $\mathrm{d}^2_{\mathrm{DCJ}}(\mathbb{S}^f, \mathbb{D}^f)$

Assign arbitrarily indices $\mathtt{a}$ and $\mathtt{b}$ to the two genes of each family in $\mathbb{D}$, obtaining $\mathbb{D}'$

All possible adjacencies in $2{\cdot}\mathbb{S}$:

For each $uv \in \Gamma(\mathbb{S})$

- the **paralogous adjacencies** are
  - either $P(uv) = \{u_{\mathtt{a}}v_{\mathtt{a}}, u_{\mathtt{b}}v_{\mathtt{b}}\}$
  - or $\widetilde{P}(uv) = \{u_{\mathtt{a}}v_{\mathtt{b}}, u_{\mathtt{b}}v_{\mathtt{a}}\}$
- and the **square** of uv is defined as $Q(uv) = P(uv) \cup \widetilde{P}(uv)$

DCJ double distance: ambiguous breakpoint graph

The **ambiguous breakpoint graph** $ABD(\mathbb{D}', 2\cdot\mathbb{S}) = (V, E)$:

1. $V = \bigcup_{\mathtt{x} \in \mathcal{G}_\star} \{\mathtt{x}_\mathtt{a}^h, \mathtt{x}_\mathtt{a}^t, \mathtt{x}_\mathtt{b}^h, \mathtt{x}_\mathtt{b}^t\} \quad \Rightarrow V = \xi(\mathbb{D}'); \quad |V| = 4n$

   there are two vertices for each extremity of each gene in $\mathcal{G}_\star$

   each vertex v has a label $\ell(v)$, that corresponds to the extremity of $\mathbb{D}'$ it represents

2. $E = E_\Gamma(\mathbb{D}') \cup E_Q(2\cdot\mathbb{S})$, where:

   - $\mathbb{D}$**-adjacency edges:** $E_\Gamma(\mathbb{D}') = \{uv : u, v \in V(\xi(\mathbb{D}')) \text{ and } \ell(u)\ell(v) \in \Gamma(\mathbb{D}')\}$
   - **Ambiguous** $\mathbb{S}$**-adjacency edges:** $E_Q(2\cdot\mathbb{S}) = \bigcup_{uv \in \Gamma(\mathbb{S})} \{uv : u, v \in V(\xi(\mathbb{D}')) \text{ and } \ell(u)\ell(v) \in Q(uv)\}$

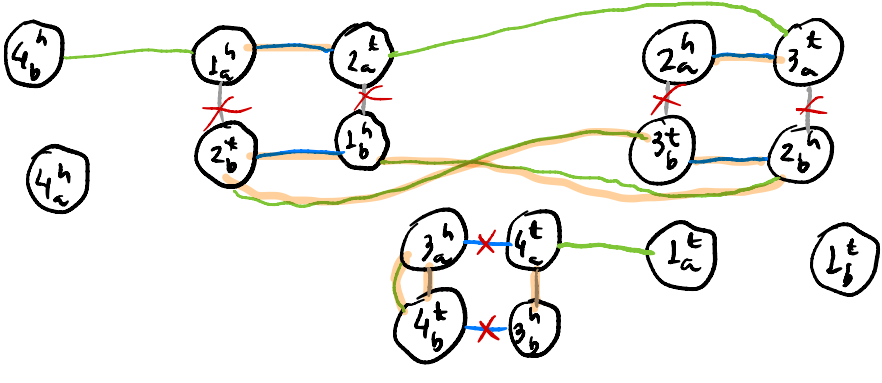
   The number of edges is $|E| = |E_\Gamma(\mathbb{D})| + 4|E_\Gamma(\mathbb{S})|$ $\begin{cases} |E_\Gamma(\mathbb{D})| \leq 2n \\ |E_\Gamma(\mathbb{S})| \leq n \end{cases}$

Ambiguous breakpoint graph

Ex: $\mathbb{D} = [\bar{4}\ 1\ \bar{4}\ \bar{3}\ 2]\ [1\ \bar{2}\ 3] \Rightarrow \begin{cases} \mathbb{D}' = [\bar{4}_a\ 1_a\ \bar{4}_b\ \bar{3}_a\ 2_a]\ [1_b\ \bar{2}_b\ 3_b] \\ \Gamma(\mathbb{D}') = \{\ 4^t_a 1^t_a,\ 1^h_a 4^h_b,\ 4^t_b 3^h_a,\ 3^t_a 2^t_a,\ 1^h_b 2^h_b,\ 2^t_b 3^t_b\ \} \end{cases}$

$\mathbb{S} = [1\ 2\ 3\ 4]$
$\Gamma(\mathbb{S}) = \{\ 1^h 2^t,\ 2^h 3^t,\ 3^h 4^t\ \}$

$\Rightarrow \begin{cases} Q(1^h 2^t) = \{1^h_a 2^t_a, 1^h_b 2^t_b, 1^h_a 2^t_b, 1^h_b 2^t_a\} \\ Q(2^h 3^t) = \{2^h_a 3^t_a, 2^h_b 3^t_b, 2^h_a 3^t_b, 2^h_b 3^t_a\} \\ Q(3^h 4^t) = \{3^h_a 4^t_a, 3^h_b 4^t_b, 3^h_a 4^t_b, 3^h_b 4^t_a\} \end{cases}$



Solution: for each **square** $Q(uv)$, fix either $P(uv)$ or $\widetilde{P}(uv)$ so that the number of cycles is maximized.

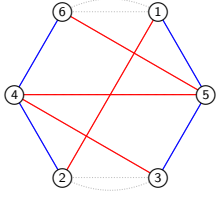
Bicolored graph of two unsigned canonical chromosomes

Each vertex of a bicolored graph has degree 0, 2 or 4:

Unsigned canonical circular chromosomes

$\widehat{\mathbb{A}} = (1\ 5\ 3\ 2\ 4\ 6)$

$\widehat{\mathbb{B}} = (1\ 2\ 3\ 4\ 5\ 6)$

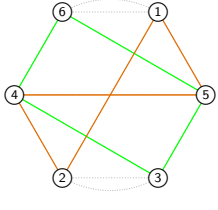


Idea:

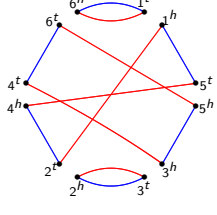
Entirely decompose a bicolored graph

into edge-disjoint alternating even cycles

One possible decomposition:



$\Rightarrow$



$\mathbb{A}_1 = (1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6)$

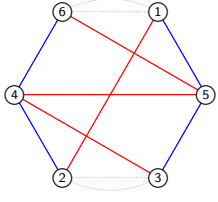
Bicolored graph of two unsigned canonical chromosomes

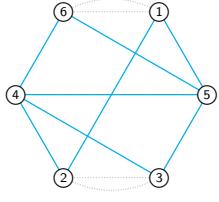
Another possible decomposition:

Unsigned canonical circular chromosomes

$\widehat{\mathbb{A}} = (1\ 5\ 3\ 2\ 4\ 6)$

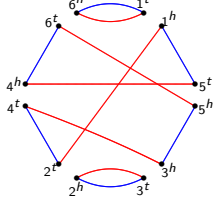
$\widehat{\mathbb{B}} = (1\ 2\ 3\ 4\ 5\ 6)$





$\Rightarrow$

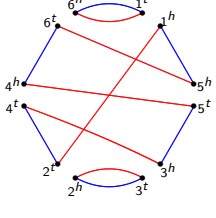
$(1\ 5\ 4\ 6\ 5\ 3\ 4\ 2)$



$\mathbb{A}_2 = (1\ 5\ \bar{3}\ \bar{2}\ 4\ 6)$

OR

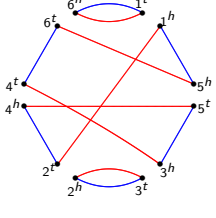
$(1\ 5\ 6\ 4\ 5\ 3\ 4\ 2)$



$\mathbb{A}_3 = (1\ \bar{5}\ \bar{3}\ \bar{2}\ 4\ 6)$

OR

$(1\ 5\ 6\ 4\ 3\ 5\ 4\ 2)$

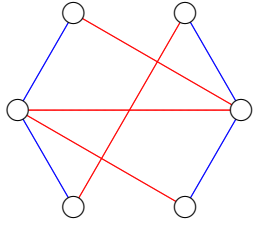


$\mathbb{A}_4 = (1\ \bar{5}\ \bar{3}\ \bar{2}\ \bar{4}\ 6)$

Bicolored graph decomposition (BGDec)

Each vertex of a bicolored graph has degree 0, 2 or 4

The number of red and of blue edges inciding in each vertex is identical



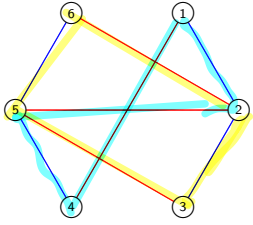
Problem:

Entirely decompose a bicolored graph
into the maximum number of edge-disjoint
alternating even cycles

$\Downarrow$

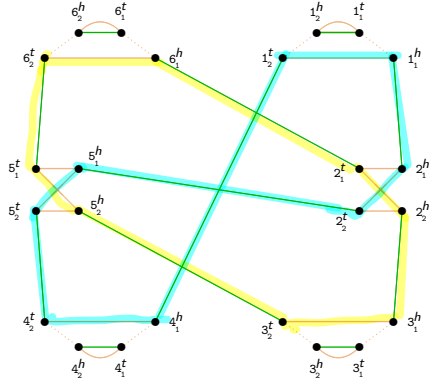
NP-hard

Reducing BGDEC to the DCJ double distance



⇒

Ambiguous breakpoint graph of $\mathbb{D}$ and $2\cdot\mathbb{S}$:



$\mathbb{D}$

$2\cdot\mathbb{S}$

DCJ median of three canonical genomes

Given three canonical genomes $\mathbb{A}$, $\mathbb{B}$, $\mathbb{C}$, find another canonical genome $\mathbb{M}$ that minimizes the sum

$$d_{DCJ}(\mathbb{M}, \mathbb{A}) + d_{DCJ}(\mathbb{M}, \mathbb{B}) + d_{DCJ}(\mathbb{M}, \mathbb{C})$$

Example:

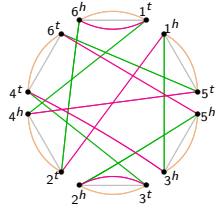
Genomes

$\mathbb{A} = (1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6)$

$\mathbb{B} = (1\ \bar{3}\ 4)\ \ (2\ \bar{5}\ 6)$

$\mathbb{C} = (1\ 2\ 3\ 4\ 5\ 6)$

Breakpoint graph of $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{C}$



Median candidate

$\mathbb{M} = \mathbb{A} = (1\ 5\ \bar{3}\ \bar{2}\ \bar{4}\ 6)$

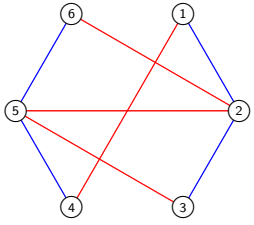
$d_{DCJ}(\mathbb{M}, \mathbb{A}) = 0$

$d_{DCJ}(\mathbb{M}, \mathbb{B}) = 6 - 2 = 4$

$d_{DCJ}(\mathbb{M}, \mathbb{C}) = 6 - 4 = 2$

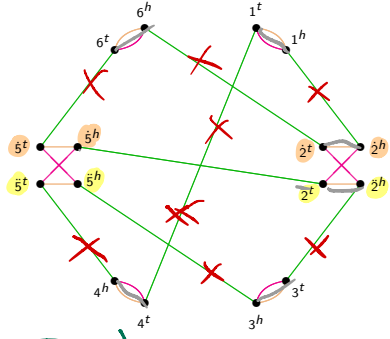
Reducing BGDEC to the DCJ median of three canonical genomes

$F = \{1, 2, 2', 3, 4, 5, 5'\}$



⇒

Breakpoint graph of $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{C}$



$\mathbb{A} = (1\ \bar{2}\ \bar{6}\ 5\ 2'\ 3\ \bar{5'}\ \bar{4})$

$\mathbb{B} = (1)\ (2\ 2')\ (3)\ (4)\ (5\ 5')\ (6)$

$\mathbb{C} = (1)\ (2)\ (2')\ (3)\ (4)\ (5)\ (5')\ (6)$

Quiz 2

1 Which of the following statements are true?

A The multi mixed/circular DCJ double distance is NP-hard, therefore the multi mixed/circular DCJ halving is also NP-hard.

B The multi linear breakpoint double distance is polynomial, therefore the multi linear breakpoint halving is also polynomial.

2 We prove that DCJ median is NP-hard...

A ... by reducing it to the bicolored graph decomposition.

B ... by reducing the bicolored graph decomposition to it.

Canonical inversion model - circular chromosomes

(Unichromosomal genomes $\equiv$ chromosomes)

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

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

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

Canonical Inversion Sorting Problem: Find a sequence of $d_{INV}(\mathbb{A}, \mathbb{B})$ inversions that transform $\mathbb{A}$ into $\mathbb{B}$.

Breakpoint diagram of canonical circular chromosomes

Let $\mathbb{A}$ and $\mathbb{B}$ be canonical circular chromosomes, with $n = |\mathcal{G}_\star|$.

The **breakpoint diagram** $BD(\mathbb{A}, \mathbb{B}) = (V, E)$ is described as follows:

1. $V = \bigcup_{\mathtt{x} \in \mathcal{G}_\star} \{\mathtt{x}^h, \mathtt{x}^t\} \quad \Rightarrow V = \xi(\mathbb{A}) = \xi(\mathbb{B}) \;\; ; \;\; |V| = 2n$

   there is a vertex for each extremity of each gene in $\mathcal{G}_\star$

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

   The vertices are drawn in one line, next to each other.

   The vertices must follow the same (circular) order of the corresponding extremities in chromosome $\mathbb{A}$, according to one of the two reading directions.

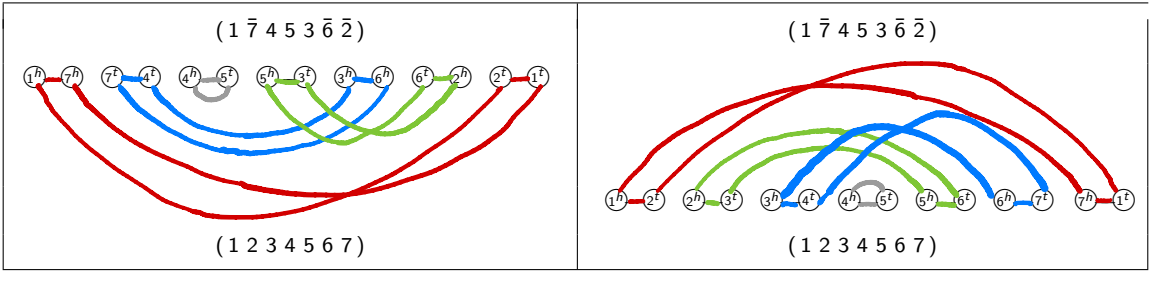
2. $E = E_\Gamma(\mathbb{A}) \cup E_\Gamma(\mathbb{B})$, where:

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

     The number of edges is $|E| = 2n$ (n adjacency edges per chromosome)

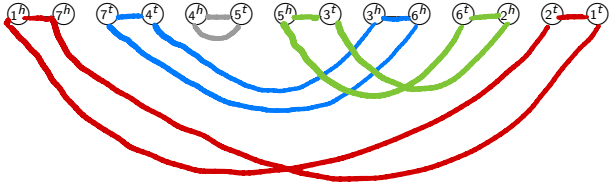
Two equivalent breakpoint diagrams

$$BD(\mathbb{A},\mathbb{B}) \;\cong\; BD(\mathbb{B},\mathbb{A})$$



Properties of the breakpoint diagram

$\mathbb{A} =$ $(1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2})$



$\mathbb{B} =$ $(1\ 2\ 3\ 4\ 5\ 6\ 7)$

$n = |\mathcal{G}_\star| = 7$

Every vertex has degree two:

$BD(\mathbb{A}, \mathbb{B})$ is a collection of (even) cycles (alternating edes in $E_\Gamma(\mathbb{A})$ and in $E_\Gamma(\mathbb{B})$)

cycle with k edges: k-cycle (always even)

$\mathcal{C}$ = set of cycles in $BD(\mathbb{A}, \mathbb{B})$

If $\mathbb{A} = \mathbb{B}$,
$RG(\mathbb{A}, \mathbb{B})$ has only 2-cycles:

$2n = 2|\mathcal{C}| \quad \Rightarrow \quad n = |\mathcal{C}|$

Otherwise, if $\mathbb{A} \neq \mathbb{B}$:

$n > |\mathcal{C}|$

References

Genome Halving under DCJ Revisited

(Julia Mixtacki)

LNCS, volume 5092, pages 276-286 (2008)

Multichromosomal median and halving problems under different genomic distances

(Eric Tannier, Chunfang Zheng and David Sankoff)