Topics of today:

1. NP-hardness of unichromosomal breakpoint median
2. Double-cut-and-join (DCJ) model
3. General DCJ halving

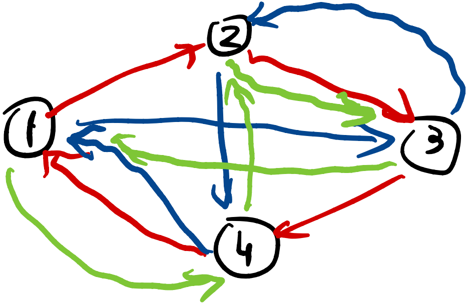
NP-hardness of unichromosomal breakpoint median

A unichromosomal circular genome $\mathbb{C}$ can be represented as a simple directed cycle graph:

Ex: $\mathbb{C} = (1\,\bar{2}\,3)$

Assume that the genes in three canonical circular genomes $\mathbb{C}_1$, $\mathbb{C}_2$ and $\mathbb{C}_3$ have the same relative orientation and represent these three genomes in the same directed cycle graph:

Ex: $\mathbb{C}_1 = (1\,2\,3\,4)$, $\mathbb{C}_2 = (2\,4\,1\,3)$, $\mathbb{C}_3 = (2\,3\,1\,4)$

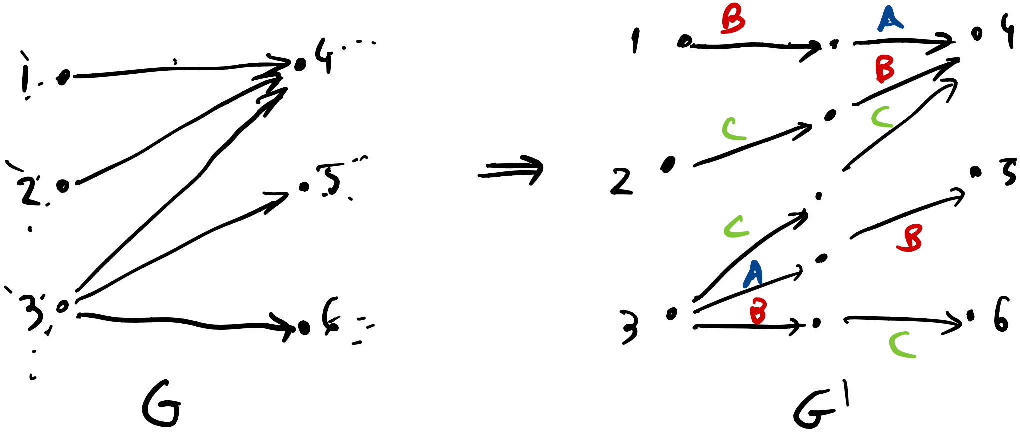


NP-hardness of unichromosomal breakpoint median

The Problem of determining whether a directed graph G has a hamiltonian cycle is NP-complete, even if G has maximum indegree and maximum outdegree equal to 3.

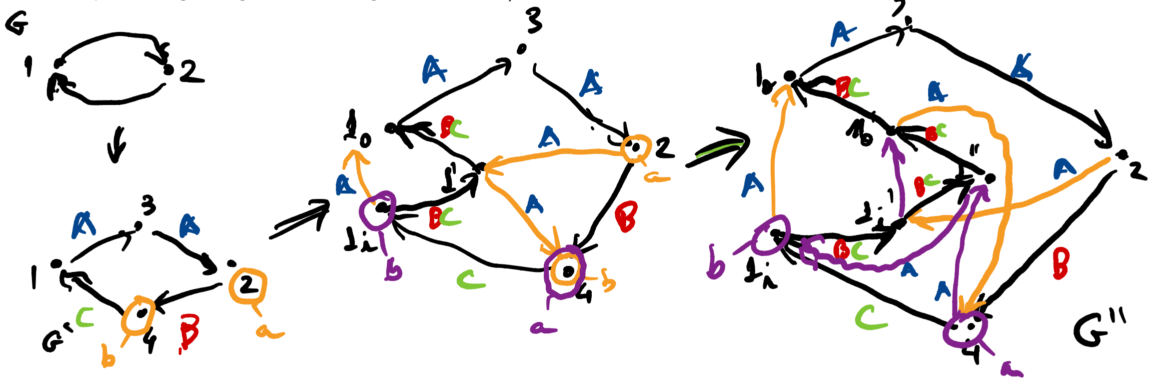
Reduction of this problem to the problem of computing a breakpoint median of three canonical circular genomes $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ that have the same relative orientation:

> We need to transform G into another directed graph G'', such that G'' is the union of three hamiltonian cycles (each one representing one input genome of the median problem)



NP-hardness of unichromosomal breakpoint median

Build a modified directed graph G'', such that G'' is the union of three hamiltonian cycles (each one representing one genome among $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$)



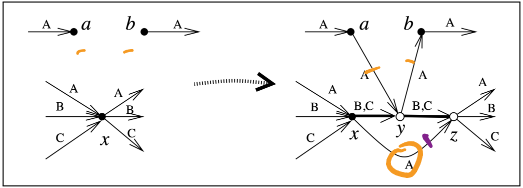
G'' has only adjacencies that occur in one or in two genomes

Let $\mathbb{M}$ be a solution to the circular breakpoint median of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$:

$\mathbb{M}$ contains all adjacencies common to two input genomes and no "new" adjacency

$\updownarrow$

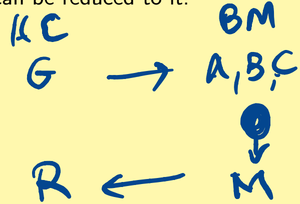
Initial graph G has an hamiltonian cycle



Quiz 1

1 Which of the following statements are true?

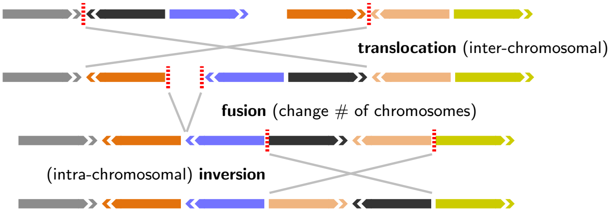
- A There is a polynomial time algorithm for solving the unichromosomal breakpoint median.
- B There cannot be a polynomial time algorithm for solving the unichromosomal breakpoint median.
- C The unichromosomal breakpoint median is NP-hard because it can be reduced to the hamiltonian cycle problem.
- D The unichromosomal breakpoint median is NP-hard because the hamiltonian cycle problem can be reduced to it.



Double-cut-and-join (DCJ) model

Double-cut-and-join (DCJ) operation: two cuts + two joins

- Cuts the genome twice and rejoins loose ends in a different way.
- Represents most large-scale genome rearrangements (inversions, translocations, fusions, fissions...)



DCJ model

DCJ operation involving two adjacencies

$xv + wz$

$xz + wv$ $\quad$ $xw + vz$

two possibilities of rejoining in a different way

Cases:

A. Each adjacency is in a distinct linear chromosome:

$[1 \; {}^{x}{}^{v} \; 2 \; 3] \quad [4 \; {}^{w}{}^{z} \; 5 \; 6]$

reciprocal translocation $\quad$ reciprocal translocation

$[1 \; {}^{x}{}^{z} \; 5 \; 6] \quad [4 \; {}^{w}{}^{v} \; 2 \; 3]$ $\quad$ reciprocal translocation $\quad$ $[1 \; {}^{x}{}^{w} \; \bar{4}] \quad [\bar{3} \; \bar{2} \; {}^{v}{}^{z} \; 5 \; 6]$

B. Both adjacencies are in the same chromosome, or one is in a circular chromosome:

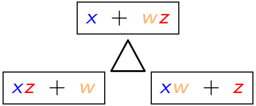
$([1 \; {}^{x}{}^{v} \; 2 \; 3 \; 4 \; {}^{z}{}^{w} \; 5 \; 6])$

inversion $\quad$ excision/ integration

$([1 \; {}^{x}{}^{z} \; \bar{4} \; \bar{3} \; \bar{2} \; {}^{v}{}^{w} \; 5 \; 6])$ $\quad$ excision/ integration $\quad$ $([1 \; {}^{x}{}^{w} \; 5 \; 6]) \quad (3 \; 4 \; {}^{z}{}^{v} \; 2)$

DCJ model

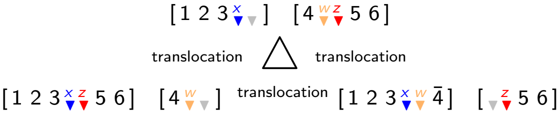
DCJ operation involving one adjacency and one telomere



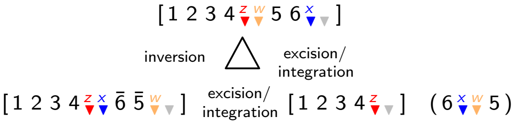
two possibilities of rejoining in a different way

Cases:

A. The adjacency and the telomere are in distinct linear chromosomes:



B. The adjacency is in the same linear chromosome, or in a circular chromosome:



DCJ model

DCJ operation involving one adjacency or two telomeres

$x + z$

$\updownarrow$

xz

one possibility of rejoining in a different way

Cases:

A. The adjacency is in a linear chromosome / the telomeres are in two distinct chromosomes:

[1 2 3 x] [z 4 5]

fusion ↓↑ fission

[1 2 3 x z 4 5] [ ]

B. The adjacency is in a circular chromosome / the telomeres are in the same chromosome:

[x 1 2 3 4 5 z]

circularization ↓↑ linearization

(2 3 4 5 z x 1) [ ]

Quiz 2

1 Which transformations can be done with a single DCJ operation?

A [1 2 3] [4 5] $\leftrightarrow$ [1 2 4 5 3]

B [1 2 3] [4 5] $\leftrightarrow$ [1 2 3 4 5]

C [1 2 3] [4 5] $\leftrightarrow$ [1 2 5] [4 3]

D [1 2 3 4 5] $\leftrightarrow$ [1 $\bar{4}$ 3 $\bar{2}$ 5]

E [1 2 3 4 5] $\leftrightarrow$ [1 2 $\bar{5}$ $\bar{4}$ $\bar{3}$]

F [1 2 3] (4 5) $\leftrightarrow$ [1 2 4 5 3]

G [1 2 3] (4 5) $\leftrightarrow$ [1 2 5 4 3]

H (1 2 3 4 5) $\leftrightarrow$ [3 4 5 1 2]

DCJ halving

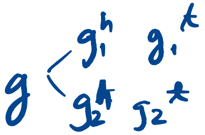
DCJ Halving Distance Problem:

Compute the minimum number of DCJ operations required to transform a (rearranged) duplicated genome $\mathbb{D}$ into a perfectly duplicated genome $2 \cdot \mathbb{H}$.

Denote by $h_{\text{DCJ}}(\mathbb{D})$ the DCJ halving distance of $\mathbb{D}$.

DCJ Halving Problem:

Find a sequence of $h_{\text{DCJ}}(\mathbb{D})$ DCJ operations that transform a (rearranged) duplicated genome $\mathbb{D}$ into a perfectly duplicated genome $2 \cdot \mathbb{H}$.



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

1. $V = \alpha(\mathbb{D}) \cup \gamma(\mathbb{D})$ (each adjacency or telomere of $\mathbb{D}$ is a vertex of $NG(\mathbb{D})$)
2. For each family $f \in \mathcal{F}(\mathbb{D})$, each pair of paralogous extremities is connected by an edge in $NG(\mathbb{D})$, i.e.:
   - there is an edge connecting the vertex u that contain f_1^h and the vertex v that contain f_2^h
   - there is an edge connecting the vertex u' that contain f_1^t and the vertex v that contain f_2^t

Note that:

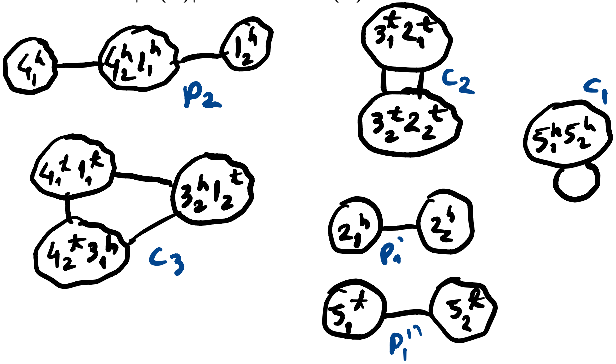
- There can be adjacencies/vertices of type $f_1^h f_2^h$ and/or $f_1^t f_2^t$ ($NG(\mathbb{D})$ can contain 1-cycles)
- Let $n = |\mathcal{F}(\mathbb{D})| = \frac{|\mathcal{G}(\mathbb{D})|}{2}$. The number of edges in $NG(\mathbb{D}) = 2n$ (two edges per element of $\mathcal{F}(\mathbb{D})$).

Natural graph of a duplicated genome

Ex: $\mathbb{D} = [\bar{4}\ 1\ \bar{4}\ \bar{3}\ 2]\quad [\bar{2}\ 3\ 1]\quad [5\ \bar{5}]$

$\alpha(\mathbb{D}) \cup \gamma(\mathbb{D}) = \{ 4_1^h,\ 4_1^t 1_1^t,\ 1_1^h 4_2^h,\ 4_2^t 3_1^h,\ 3_1^t 2_1^t,\ 2_1^h,\ 2_2^h,\ 2_2^t 3_2^t,\ 3_2^h 1_2^t,\ 1_2^h,\ 5_1^t,\ 5_1^h 5_2^h,\ 5_2^t \}$

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



Every vertex has degree one or two:
$NG(\mathbb{D})$ 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

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

$= \{c_2\}$
$= \{p_2\}$
$= \{c_1, c_3\}$
$= \{p_1', p_1''\}$

$|\mathcal{C}_o| + |\mathcal{P}_o|$ is even (NG has $2n$ edges)

$|\mathcal{P}_e| + |\mathcal{P}_o| = \kappa(\mathbb{D})$

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

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

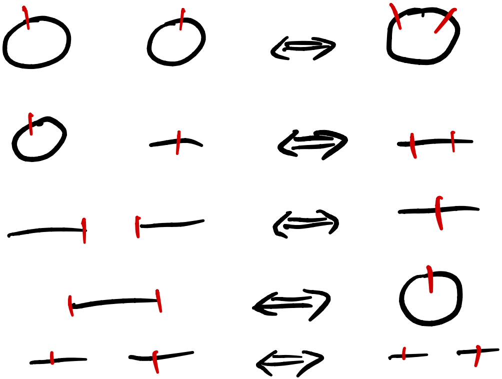
Otherwise, if a duplicated genome $\mathbb{D}$
is not perfectly duplicated:

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

Types of DCJ operation

Let a DCJ operation transform a duplicated genome $\mathbb{D}_1$ into another duplicated genome $\mathbb{D}_2$:

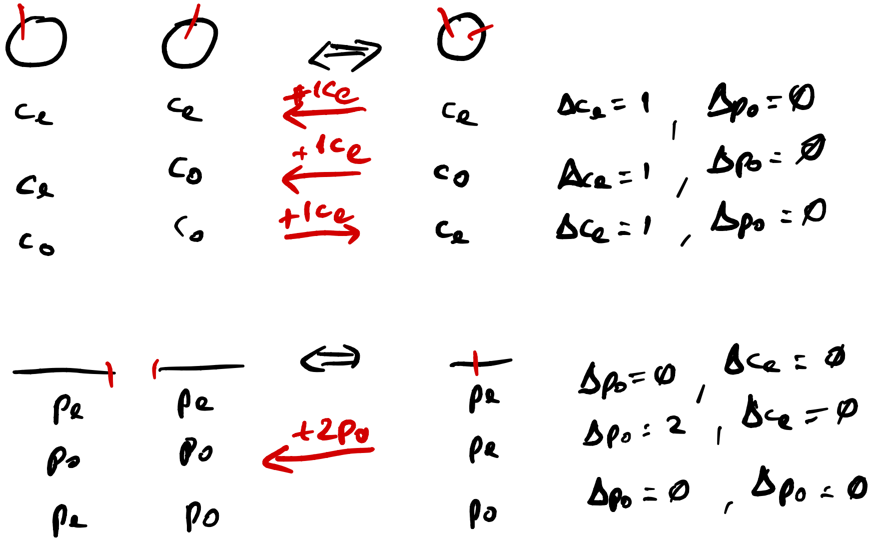
$$\left.\begin{array}{l} m_1 : \#\text{ of components in } NG(\mathbb{D}_1) \\ m_2 : \#\text{ of components in } NG(\mathbb{D}_2) \end{array}\right\} \quad 0 \leq |m_2 - m_1| \leq 1$$



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

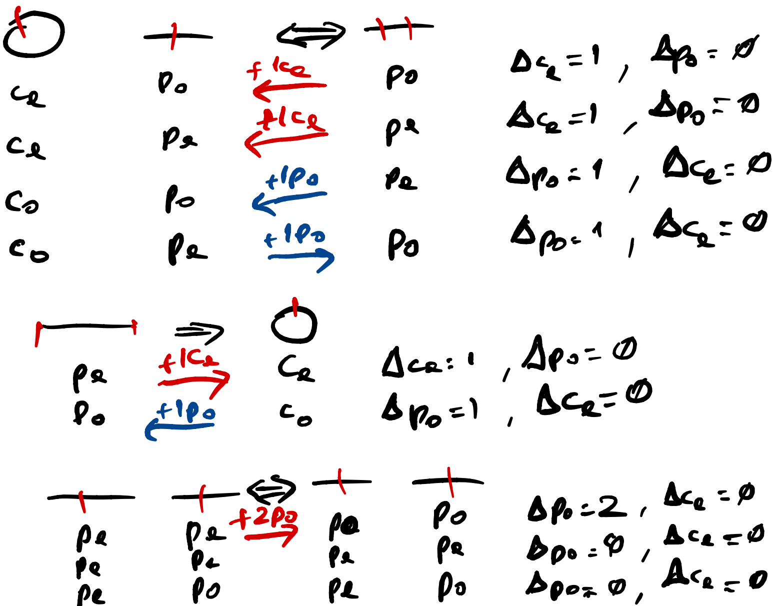
Types of DCJ operation

Goal: increase the number of even cycles ($|\mathcal{C}_e|$) and/or 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 & Distance

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

DCJ +1

+2

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 $\mathbb{D}$, it is possible to find an optimal DCJ operation at each sorting step. Therefore:

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

DCJ Halving

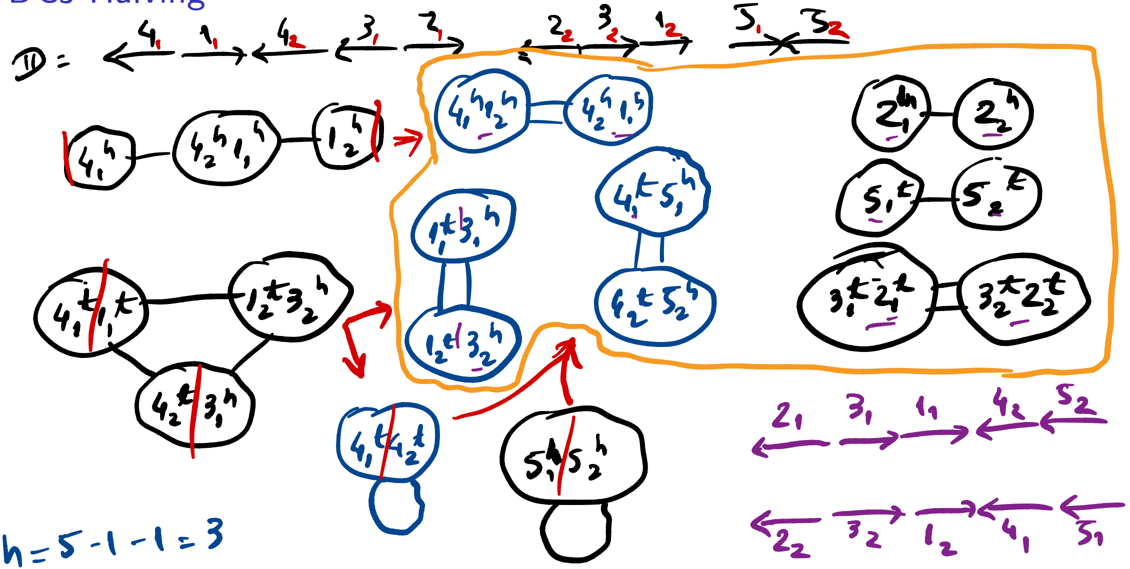
Given a duplicated genome $\mathbb{D}$,

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

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

1. For $i = 1$ to h :
   - Find and apply one optimal DCJ operation.
2. NG is now a simple collection of 2-cycles and 1-paths.
   Reconstruct the perfectly duplicated genome $2{\cdot}\mathbb{H}$ from NG.

DCJ Halving



$h = 5 - 1 - 1 = 3$

Quiz 3

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.
- C Breaking an even path into two odd paths is always optimal.
- D Breaking an even cycle into two cycles is always optimal.
- E Recombining two even paths into two odd paths is always optimal.

can be optimal if $|P_0|$ was odd

optimal only when the two new cycles are even

References

The complexity of the breakpoint median problem

(David Bryant)

Tech. Rep. CRM-2579, Centre de recherches mathématiques, Université de Montréal, 1998

Genome Halving under DCJ Revisited

(Julia Mixtacki)

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