Topics of today:

1. Breakpoint graph of two genomes
2. SCJ circular median
3. Breakpoint median and halving
4. NP-hardness of unichromosomal breakpoint median
5. Large-scale rearrangements and the DCJ operation

Breakpoint graph of canonical genomes

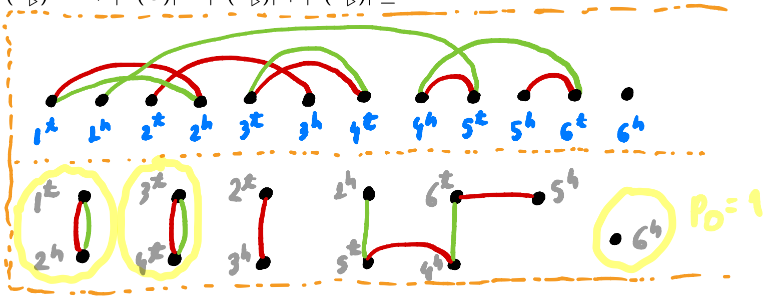
Genomes $\mathbb{A}^f_\triangleright$ and $\mathbb{B}^f_\triangleright$ are canonical, with $\mathcal{F}_\star = \mathcal{G}_\star = \{1, 2, \dots, n\}$

Breakpoint graph of $\mathbb{A}^f_\triangleright$ and $\mathbb{B}^f_\triangleright$ → $G = BG(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright)$:

1. Set of vertices $V(G) = \bigcup_{\mathtt{X} \in \mathcal{G}_\star} \{\mathtt{X}^h, \mathtt{X}^t\}$ $\Rightarrow |V(G)| = 2n$
2. Set of edges $E(G) = \Gamma(\mathbb{A}^f_\triangleright) \cup \Gamma(\mathbb{B}^f_\triangleright)$ $\Rightarrow |E(G)| = |\Gamma(\mathbb{A}^f_\triangleright)| + |\Gamma(\mathbb{B}^f_\triangleright)| \leq 2n$

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

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



1^t 1^h 2^t 2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h
1^t
2^h
3^t
4^t
2^t
3^h
1^h
5^t
6^t
5^h
4^h
6^h
$p_0 = 1$

$c_2 = 2$

Each vertex has degree $\overset{0,}{1}$ or 2: collection of p paths and c (even) cycles ($p = \kappa(\mathbb{A}^f_\triangleright) + \kappa(\mathbb{B}^f_\triangleright)$)

length of a component: number of edges alternating

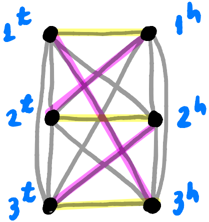
$$d_{\mathrm{BP}}(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) = n - c_2 - \frac{p_0}{2} \begin{cases} c_2 = \text{number of 2-cycles in } BG(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) \text{ (common adjacencies)} \\ p_0 = \text{number of 0-paths in } BG(\mathbb{A}^f_\triangleright, \mathbb{B}^f_\triangleright) \text{ (common telomeres)} \end{cases}$$

Complete graph of a set of genes

Complete graph $\mathfrak{G}$ of $\mathcal{A} = \{1, 2, \ldots, n\}$:

Set of vertices $V(\mathfrak{G}) = \bigcup_{\mathtt{X} \in \mathcal{A}} \{\mathtt{X}^h, \mathtt{X}^t\} \quad \Rightarrow |V(\mathfrak{G})| = 2n$

Ex : $n = 3$



$C_{🟡} = (1)\ (2)\ (3)$

$C_{🟣} = (1\ \ 2\ \ 3)$

Set of non-incident edges covering all vertices

A perfect matching M in $\mathfrak{G}$ corresponds to $|M| = n$ adjacencies and, consequently, defines a circular singular genome $\mathbb{C}$, with $\Gamma(\mathbb{C}) = M$.

SCJ median of k canonical genomes

Given k canonical genomes $\mathbb{C}_1^f$, $\mathbb{C}_2^f$, $\ldots \mathbb{C}_k^f$, find another canonical genome $\mathbb{M}^f$ that minimizes the sum:

$$\begin{aligned} \mathsf{s}_{\mathrm{SCJ}}(\mathbb{M}^f) &= \mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathbb{C}_1^f) + \mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathbb{C}_2^f) + \ldots + \mathsf{d}_{\mathrm{SCJ}}(\mathbb{M}^f, \mathbb{C}_k^f) \\ &= |\Gamma(\mathbb{C}_1^f)| + |\Gamma(\mathbb{C}_2^f)| + \ldots + |\Gamma(\mathbb{C}_k^f)| \quad + \quad \omega(\mathbb{M}^f) \end{aligned}$$

For computing the median, we need to minimize:

$$\omega(\mathbb{M}^f) \;=\; \sum_{xy \in \Gamma(\mathbb{M}^f)} \omega(xy)$$

where $\omega(xy) = k - 2 \cdot \phi(xy, \mathbb{C}_{1..k}^f) \in \{-k, -k+2, \ldots, +k-2, +k\}$.

Solution: take only the adjacencies with negative weight

SCJ median of k canonical circular genomes

Canonical genomes $\mathbb{C}^f_1, \mathbb{C}^f_2, \ldots \mathbb{C}^f_k$ are circular:

$$\begin{aligned} s_{\text{SCJ}}(\mathbb{M}^f) &= |\Gamma(\mathbb{C}^f_1)| + |\Gamma(\mathbb{C}^f_2)| + \ldots + |\Gamma(\mathbb{C}^f_k)| \quad + \quad \omega(\mathbb{M}^f) \\ &= \underset{kn}{\cancel{3n}} \quad + \quad \omega(\mathbb{M}^f) \end{aligned}$$

Again, we need to minimize $\omega(\mathbb{M}^f) \;=\; \sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \omega(xy)$, where $\omega(xy) = k - 2\cdot\phi(xy, \mathbb{C}^f_{1..k})$, but since the median is required to be circular, it may not be possible to take only the adjacencies that have a negative weight

Solution:

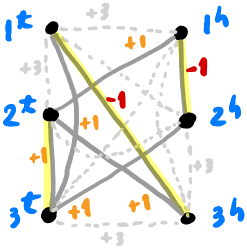
1. Build the complete graph $\mathfrak{G}$ of $\mathcal{G}_\star$
2. Assign weights to each edge xy of $\mathfrak{G}$: $\omega(xy) = k - 2\cdot\phi(xy, \mathbb{C}^f_{1..k})$.

$$\omega(xy) \in \{\underset{(3)}{-3},\ \underset{(2)}{-1},\ \underset{(1)}{+1},\ \underset{(0)}{+3}\}$$

Ex: $C_1 = (1\ 2\ 3)$

$C_2 = (3\ 2\ \bar{1})$

$C_3 = (3\ 1\ \bar{2})$



$M = (1\ \bar{2}\ 3)$

Perfect matching M in $\mathfrak{G} \Leftrightarrow$ Circular genome $\mathbb{M}^f$; with $\omega(M) = \omega(\mathbb{M}^f)$

A perfect matching M_{MIN} with **minimum weight** gives a circular SCJ median $\mathbb{M}^f_{min}$ with **minimum weight**

Breakpoint median of three canonical circular genomes

Given canonical circular genomes $\mathbb{C}^f_1$, $\mathbb{C}^f_2$ and $\mathbb{C}^f_3$, find a canonical circular genome $\mathbb{M}^f_\triangleright$ that minimizes the sum:

$$\begin{aligned}
\mathrm{s}_{\mathrm{BP}}(\mathbb{M}^f_\triangleright) &= \mathrm{d}_{\mathrm{BP}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_1) + \mathrm{d}_{\mathrm{BP}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_2) + \mathrm{d}_{\mathrm{BP}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_k) \\
&= n - \textstyle\sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \phi(xy, \mathbb{C}^f_1) + n - \sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \phi(xy, \mathbb{C}^f_2) + n - \sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \phi(xy, \mathbb{C}^f_3) \\
&= 3n - \textstyle\sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \phi(xy, \mathbb{C}^f_{1..3}) \\
&= 3n - \omega'(\mathbb{M}^f_\triangleright)
\end{aligned}$$

Here we need to maximize $\omega'(\mathbb{M}^f_\triangleright) = \sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \omega'(xy)$, where $\omega'(xy) = \phi(xy, \mathbb{C}^f_{1..3})$.

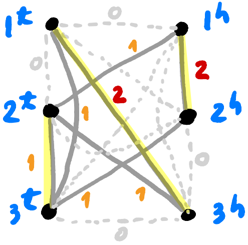
1. Build the complete graph $\mathfrak{G}$ of $\mathcal{G}_\star$
2. Assign weights to each edge xy of $\mathfrak{G}$: $\omega'(xy) = \phi(xy, \mathbb{C}^f_{1..k})$.

$\omega(xy) \in \{0, 1, 2, 3\}$

Ex: $C_1 = (1\ 2\ 3)$

$C_2 = (3\ 2\ \bar{1})$

$C_3 = (3\ 1\ \bar{2})$



$M = (1\ \bar{2}\ 3)$

Perfect matching M in $\mathfrak{G}$ ⇔ Circular genome $\mathbb{M}^f$; with $\omega'(M) = \omega'(\mathbb{M}^f)$

A perfect matching M_{MIN} with **maximum weight** gives a circular BP median $\mathbb{M}^f_{min}$ with **maximum weight**

Breakpoint median of three canonical genomes

Given canonical genomes $\mathbb{C}^f_1$, $\mathbb{C}^f_2$ and $\mathbb{C}^f_3$, find a canonical genome $\mathbb{M}^f_\triangleright$ that minimizes the sum:

$$\begin{aligned}
\mathrm{s}_{\mathrm{BP}}(\mathbb{M}^f_\triangleright) \quad &= \quad \mathrm{d}_{\mathrm{BP}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_1) \;+\; \mathrm{d}_{\mathrm{BP}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_2) \;+\; \mathrm{d}_{\mathrm{BP}}(\mathbb{M}^f_\triangleright, \mathbb{C}^f_k) \\
&= \quad n - \textstyle\sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \phi(xy, \mathbb{C}^f_1) - \textstyle\sum_{x\in\Theta(\mathbb{M}^f_\triangleright)} \frac{\phi(x,\mathbb{C}^f_1)}{2} \;+\; n - \textstyle\sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \phi(xy, \mathbb{C}^f_2) \\
&\qquad - \textstyle\sum_{x\in\Theta(\mathbb{M}^f_\triangleright)} \frac{\phi(x,\mathbb{C}^f_2)}{2} \;+\; n - \textstyle\sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \phi(xy, \mathbb{C}^f_3) - \textstyle\sum_{x\in\Theta(\mathbb{M}^f_\triangleright)} \frac{\phi(x,\mathbb{C}^f_3)}{2} \\
&= \quad 3n \;-\; \textstyle\sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \phi(xy, \mathbb{C}^f_{1..3}) \;-\; \textstyle\sum_{x\in\Theta(\mathbb{M}^f_\triangleright)} \frac{\phi(x,\mathbb{C}^f_{1..3})}{2}
\end{aligned}$$

Here $\omega'(\mathbb{M}^f_\triangleright) \;=\; \sum_{xy\in\Gamma(\mathbb{M}^f_\triangleright)} \omega'(xy) + \sum_{x\in\Theta(\mathbb{M}^f_\triangleright)} \omega'(x)$, where $\omega'(xy) = \phi(xy, \mathbb{C}^f_{1..3})$ and $\omega'(x) = \frac{\phi(x,\mathbb{C}^f_{1..3})}{2}$.

1. Build the complete graph $\mathfrak{G}$
2. Assign weights to each edge xy of $\mathfrak{G}$: $\omega'(xy) = \phi(xy, \mathbb{C}^f_{1..k})$.
3. Build the complete graph $\mathfrak{G}_t$ with vertices $V(\mathfrak{G}_t) = \bigcup\limits_{\mathtt{X}\in\mathcal{G}_\star} \{t_{\mathtt{X}^h}, t_{\mathtt{X}^t}\}$
4. Assign weight 0 to each edge of $\mathfrak{G}_t$
5. Add one edge connecting each vertex x in $\mathfrak{G}$ to the corresponding vertex t_x in $\mathfrak{G}_t$, with weight $\omega'(xt_x) = \frac{\phi(x,\mathbb{C}^f_{1..k})}{2}$

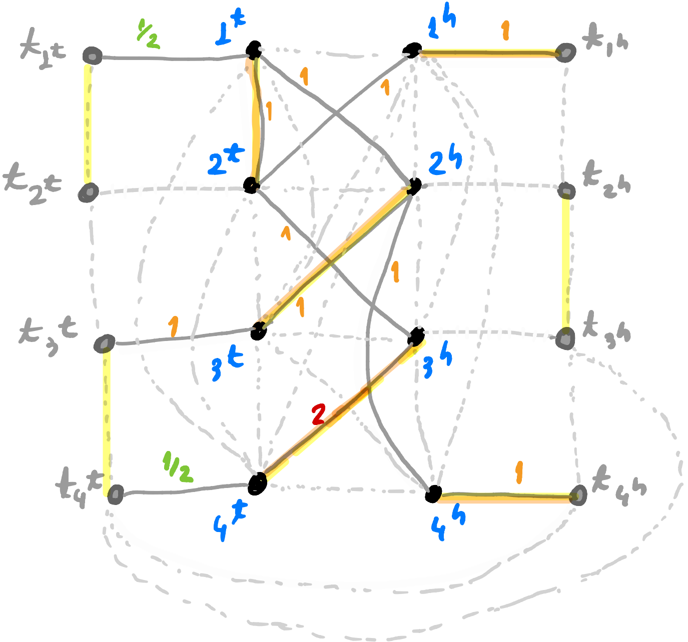
Perfect matching M in $\mathfrak{G} + \mathfrak{G}_t \Leftrightarrow$ Genome $\mathbb{M}^f$; with $\omega'(M) = \omega'(\mathbb{M}^f)$

A matching M_{MIN} with **maximum weight** gives a BP median $\mathbb{M}^f_{min}$ with **maximum weight**

$L_1 = [1\ \ 2\ \ 3\ \ 4]$

$L_2 = [3\ \ 2\ \ 1]\ [4]$

$L_3 = [\bar{1}\ \ 2\ \ \bar{4}\ \ \bar{3}]$



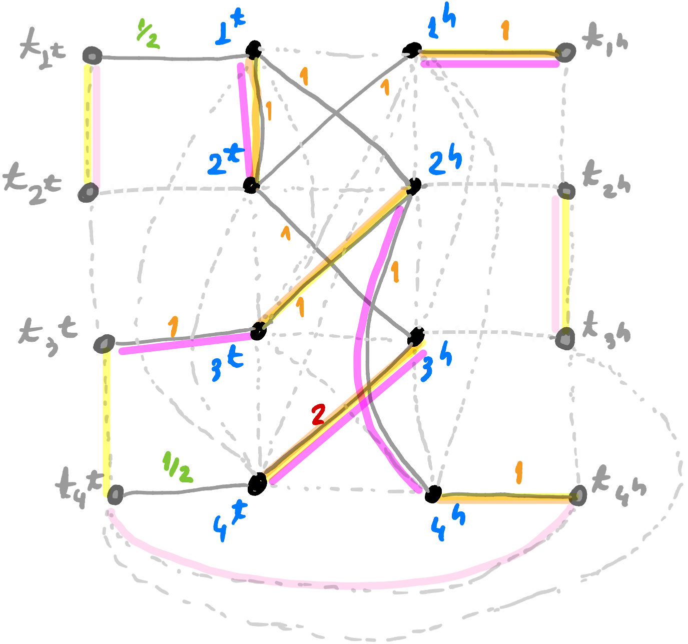
$M = [\bar{1}\ \ 2\ \ 3\ \ 4]$

$w(M) = 6$

$L_1 = [1 \;\; 2 \;\; 3 \;\; 4]$

$L_2 = [3 \;\; 2 \;\; 1] \; [4]$

$L_3 = [\bar{1} \;\; 2 \;\; \bar{4} \;\; \bar{3}]$



$M_2 = [\, 3 \;\; 4 \;\; \bar{2} \;\; 1 \,]$

$w(M_2) = 6$

The breakpoint halving can

be computed in a similar way

Quiz 1

1 Which of the following statements are true?

A The breakpoint median can only be computed for circular genomes.

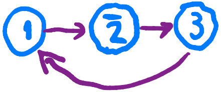
B The circular SCJ median is equivalent to the circular breakpoint median of three canonical circular genomes.

C The problem of computing a circular breakpoint halving of a circular duplicated genome is polynomial.

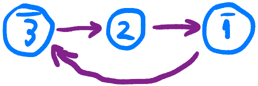
NP-hardness of unichromosomal breakpoint median

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

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

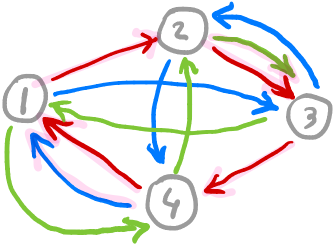


OR



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

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



Every vertex has indegree =
outdegree = 3

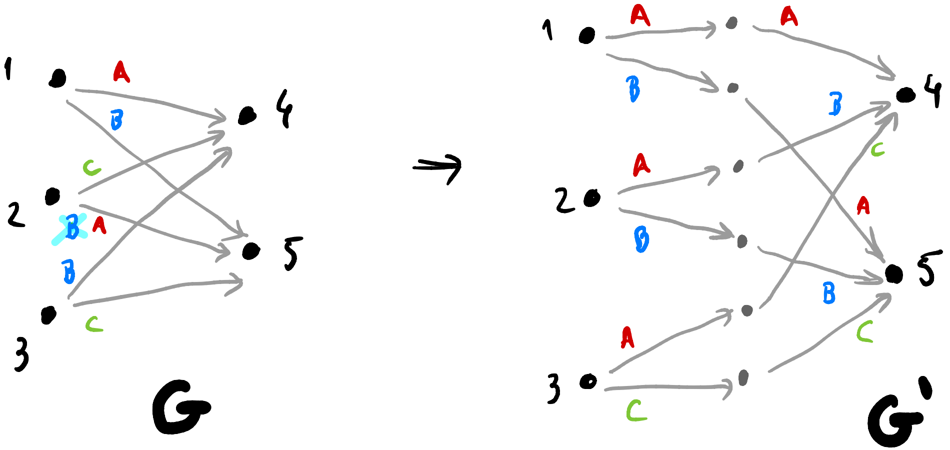
M:(1 2 3 4)

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.

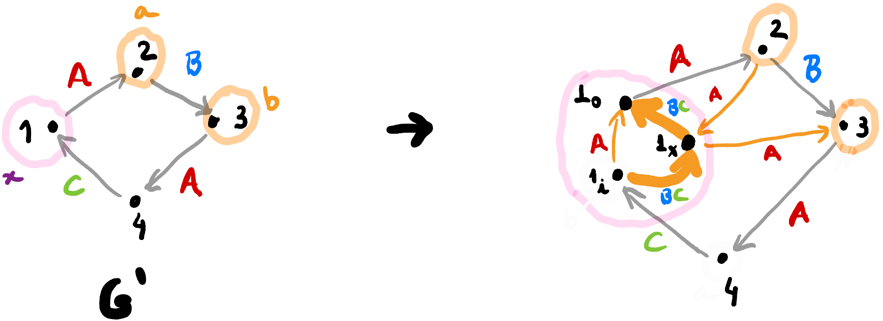
Reduction of this problem to the problem of computing a breakpoint median of three canonical circular genomes $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ that have the same relative orientation:

> 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 $\mathbf{C}$)



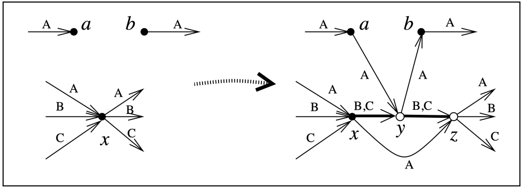
G'' has only adjacencies that occur in one or in two genomes

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

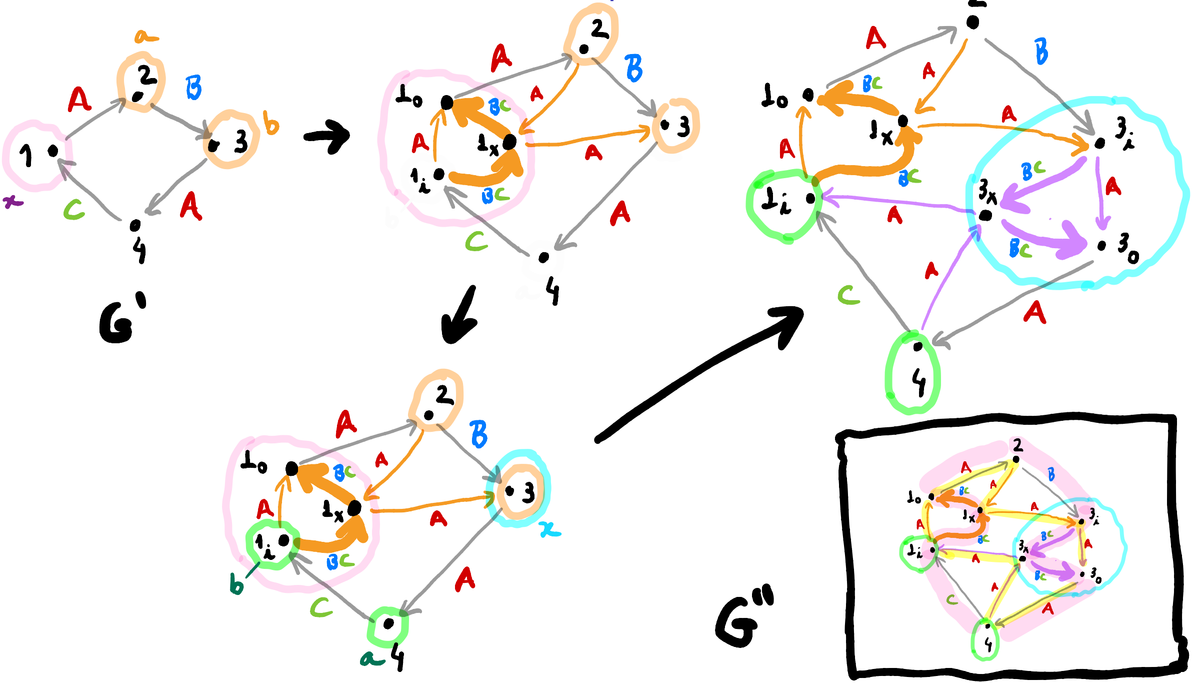
$\mathbb{M}$ contains all adjacencies common to two input genomes and no "new" adjacency

$\updownarrow$

Initial graph G has an hamiltonian cycle



NP-hardness of unichromosomal breakpoint median



Quiz 2

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.

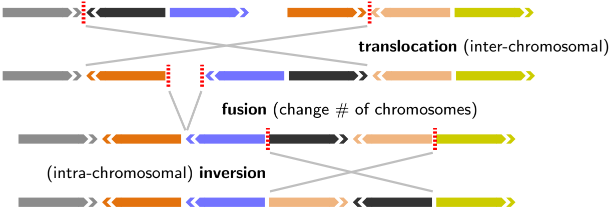
C The unichromosomal breakpoint median is NP-hard because it can be reduced to the hamiltonian cycle problem.

D The unichromosomal breakpoint median is NP-hard because the hamiltonian cycle problem can be reduced to it.

Double-cut-and-join (DCJ) model

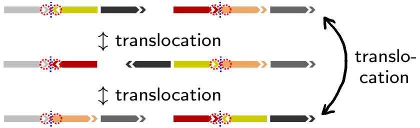
Double-cut-and-join (DCJ) operation: two cuts + two joins

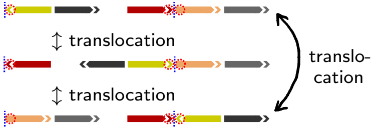
- Cuts the genome twice and rejoins loose ends in a different way.
- Represents most large-scale genome rearrangements (inversions, translocations, fusions, fissions...)


translocation (inter-chromosomal)
fusion (change # of chromosomes)
(intra-chromosomal) **inversion**

The double-cut-and-join (DCJ) operation

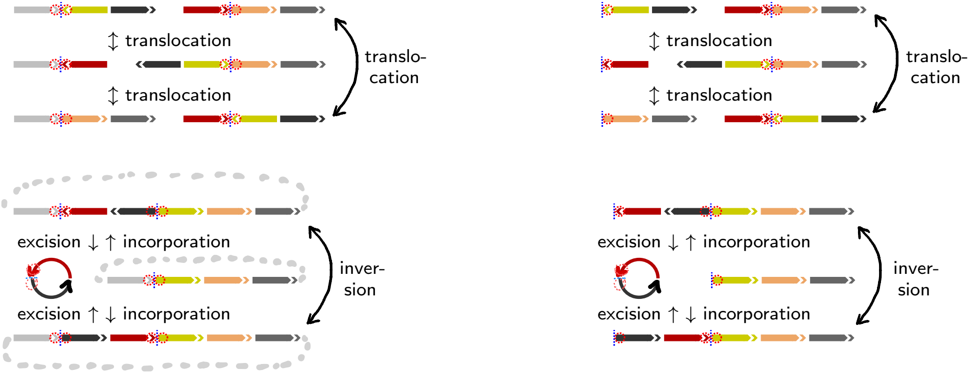
Cuts the genome in (at most) 2 positions and rejoins the open ends in a distinct way





The double-cut-and-join (DCJ) operation

Cuts the genome in (at most) 2 positions and rejoins the open ends in a distinct way



↕ translocation

translocation

↕ translocation

↕ translocation

translocation

↕ translocation

excision ↓ ↑ incorporation

excision ↑ ↓ incorporation

inversion

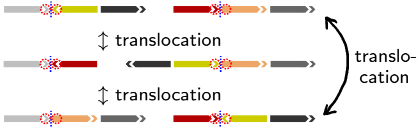
excision ↓ ↑ incorporation

excision ↑ ↓ incorporation

inversion

The double-cut-and-join (DCJ) operation

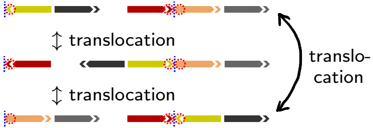
Cuts the genome in (at most) 2 positions and rejoins the open ends in a distinct way



↕ translocation

↕ translocation

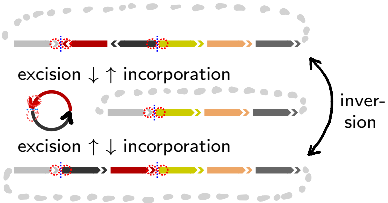
translo-cation



↕ translocation

↕ translocation

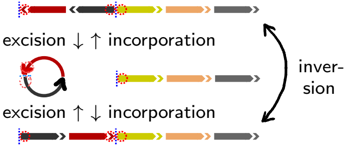
translo-cation



excision ↓ ↑ incorporation

excision ↑ ↓ incorporation

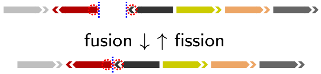
inver-sion



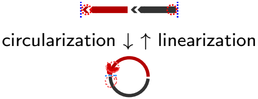
excision ↓ ↑ incorporation

excision ↑ ↓ incorporation

inver-sion



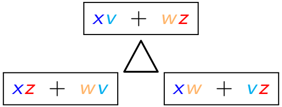
fusion ↓ ↑ fission



circularization ↓ ↑ linearization

DCJ model

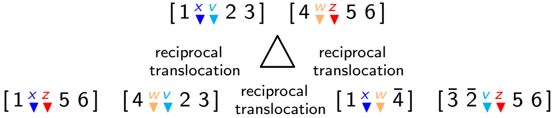
DCJ operation
involving
two adjacencies



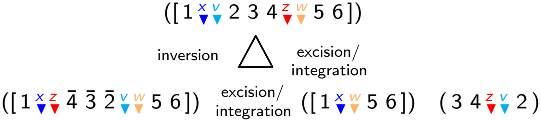
two possibilities
of rejoining
in a different way

Cases:

A. Each adjacency is in a distinct linear chromosome:

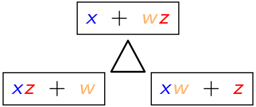


B. Both adjacencies are in the same chromosome, or one is in a circular chromosome:



DCJ model

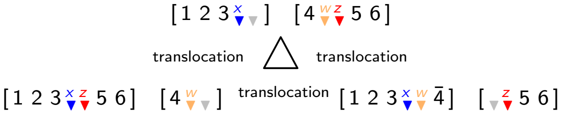
DCJ operation involving one adjacency and one telomere



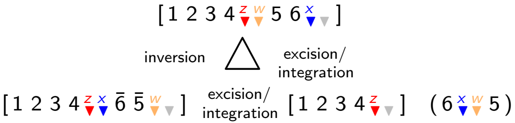
two possibilities of rejoining in a different way

Cases:

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

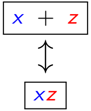


B. The adjacency is in the same linear chromosome, or in a circular chromosome:



DCJ model

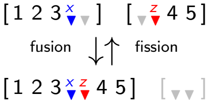
DCJ operation involving one adjacency or two telomeres



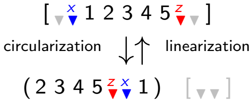
one possibility of rejoining in a different way

Cases:

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



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



Quiz 3

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

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

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

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

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

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

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

References

Multichromosomal median and halving problems under different genomic distances

(Eric Tannier, Chunfang Zheng and David Sankoff)

BMC Bioinformatics volume 10, Article number: 120 (2009)

SCJ: A Breakpoint-Like Distance that Simplifies Several Rearrangement Problems

(Pedro Feijão and João Meidanis)

TCBB volume 8 Number: 5 (2011)

The complexity of the breakpoint median problem

(David Bryant)

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