

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

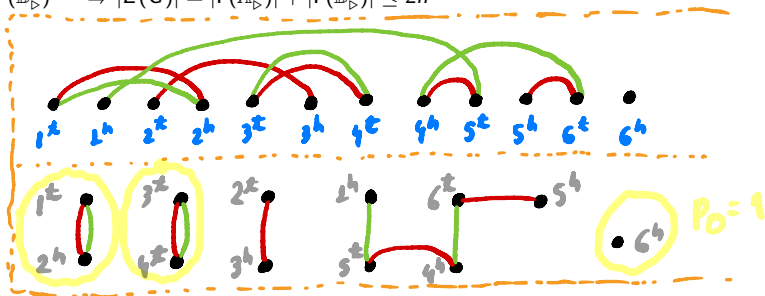
Breakpoint graph of $\mathbb{A}_\triangleright^f$ and $\mathbb{B}_\triangleright^f \rightarrow G = BG(\mathbb{A}_\triangleright^f, \mathbb{B}_\triangleright^f)$:

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

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

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

$c_2 = 2$



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

length of a component: number of edges alternating

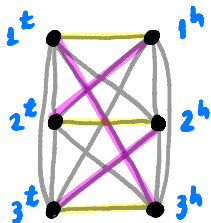
$$d_{BP}(\mathbb{A}_\triangleright^f, \mathbb{B}_\triangleright^f) = n - c_2 - \frac{p_0}{2} \begin{cases} c_2 = \text{number of 2-cycles in } BG(\mathbb{A}_\triangleright^f, \mathbb{B}_\triangleright^f) \text{ (common adjacencies)} \\ p_0 = \text{number of 0-paths in } BG(\mathbb{A}_\triangleright^f,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

Complete graph of a set of genes

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

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

$$\text{Ex: } n=3$$



$$C_{\bullet} = (1) (2) (3)$$

$$C_{\bullet} = (1 \ 2 \ 3)$$

→ Set of non-incident edges covering all vertices

A **perfect matching** M in \mathcal{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 $C_1^f, C_2^f, \dots, C_k^f$, find another canonical genome M^f that minimizes the sum:

$$\begin{aligned} s_{\text{SCJ}}(M^f) &= d_{\text{SCJ}}(M^f, C_1^f) + d_{\text{SCJ}}(M^f, C_2^f) + \dots + d_{\text{SCJ}}(M^f, C_k^f) \\ &= |\Gamma(C_1^f)| + |\Gamma(C_2^f)| + \dots + |\Gamma(C_k^f)| + \omega(M^f) \end{aligned}$$

For computing the median, we need to minimize:

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

where $\omega(xy) = k - 2 \cdot \phi(xy, C_{1..k}^f) \in \{-k, -k+2, \dots, +k-2, +k\}$.

Solution: take only the adjacencies with negative weight

SCJ median of k canonical circular genomes

Canonical genomes $C_1^f, C_2^f, \dots, C_k^f$ are circular:

$$\begin{aligned} s_{\text{SCJ}}(M^f) &= |\Gamma(C_1^f)| + |\Gamma(C_2^f)| + \dots + |\Gamma(C_k^f)| + \omega(M^f) \\ &= 3n + \omega(M^f) \end{aligned}$$

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

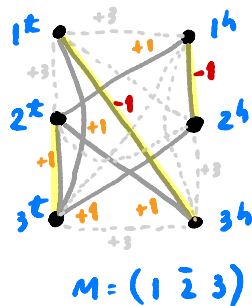
Solution:

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

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

(3) (2) (1) (0)

Ex: $C_1 = (1 \ 2 \ 3)$
 $C_2 = (3 \ 2 \ \bar{1})$
 $C_3 = (3 \cdot 1 \cdot \bar{2})$



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

A perfect matching M_{MIN} with **minimum weight** gives a circular SCJ median M_{min}^f with **minimum weight**

Breakpoint median of three canonical circular genomes

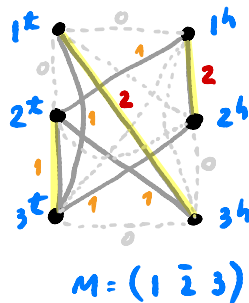
Given canonical circular genomes C_1^f , C_2^f and C_3^f , find a canonical circular genome M_{\triangleright}^f that minimizes the sum:

$$\begin{aligned}
 s_{BP}(M_{\triangleright}^f) &= d_{BP}(M_{\triangleright}^f, C_1^f) + d_{BP}(M_{\triangleright}^f, C_2^f) + d_{BP}(M_{\triangleright}^f, C_3^f) \\
 &= n - \sum_{xy \in \Gamma(M_{\triangleright}^f)} \phi(xy, C_1^f) + n - \sum_{xy \in \Gamma(M_{\triangleright}^f)} \phi(xy, C_2^f) + n - \sum_{xy \in \Gamma(M_{\triangleright}^f)} \phi(xy, C_3^f) \\
 &= 3n - \sum_{xy \in \Gamma(M_{\triangleright}^f)} \phi(xy, C_{1..3}^f) \\
 &= 3n - \omega'(M_{\triangleright}^f)
 \end{aligned}$$

Here we need to maximize $\omega'(M_{\triangleright}^f) = \sum_{xy \in \Gamma(M_{\triangleright}^f)} \omega'(xy)$, where $\omega'(xy) = \phi(xy, C_{1..3}^f)$.

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

$$\omega(xy) \in \{0, 1, 2, 3\} \quad \text{Ex: } C_1 = (1 \ 2 \ 3) \\
 C_2 = (3 \ 2 \ \bar{1}) \\
 C_3 = (3 \cdot 1 \cdot \bar{2})$$



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

A perfect matching M_{MIN} with **maximum weight** gives a circular BP median M_{min}^f with **maximum weight**

Breakpoint median of three canonical genomes

Given canonical genomes C_1^f , C_2^f and C_3^f , find a canonical genome M_{\triangleright}^f that minimizes the sum:

$$\begin{aligned}
 s_{BP}(M_{\triangleright}^f) &= d_{BP}(M_{\triangleright}^f, C_1^f) + d_{BP}(M_{\triangleright}^f, C_2^f) + d_{BP}(M_{\triangleright}^f, C_3^f) \\
 &= n - \sum_{xy \in \Gamma(M_{\triangleright}^f)} \phi(xy, C_1^f) - \sum_{x \in \Theta(M_{\triangleright}^f)} \frac{\phi(x, C_1^f)}{2} + n - \sum_{xy \in \Gamma(M_{\triangleright}^f)} \phi(xy, C_2^f) \\
 &\quad - \sum_{x \in \Theta(M_{\triangleright}^f)} \frac{\phi(x, C_2^f)}{2} + n - \sum_{xy \in \Gamma(M_{\triangleright}^f)} \phi(xy, C_3^f) - \sum_{x \in \Theta(M_{\triangleright}^f)} \frac{\phi(x, C_3^f)}{2} \\
 &= 3n - \sum_{xy \in \Gamma(M_{\triangleright}^f)} \phi(xy, C_{1..3}^f) - \sum_{x \in \Theta(M_{\triangleright}^f)} \frac{\phi(x, C_{1..3}^f)}{2}
 \end{aligned}$$

Here $\omega'(M_{\triangleright}^f) = \sum_{xy \in \Gamma(M_{\triangleright}^f)} \omega'(xy) + \sum_{x \in \Theta(M_{\triangleright}^f)} \omega'(x)$, where $\omega'(xy) = \phi(xy, C_{1..3}^f)$ and $\omega'(x) = \frac{\phi(x, C_{1..3}^f)}{2}$.

1. Build the complete graph \mathfrak{G}
2. Assign weights to each edge xy of \mathfrak{G} : $\omega'(xy) = \phi(xy, C_{1..k}^f)$.
3. Build the complete graph \mathfrak{G}_t with vertices $V(\mathfrak{G}_t) = \bigcup_{x \in \mathcal{G}_*} \{t_{x^h}, t_{x^t}\}$
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 $\omega'(xt_x) = \frac{\phi(x, C_{1..k}^f)}{2}$

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

A matching M_{MIN} with **maximum weight** gives a BP median M_{min}^f with **maximum weight**

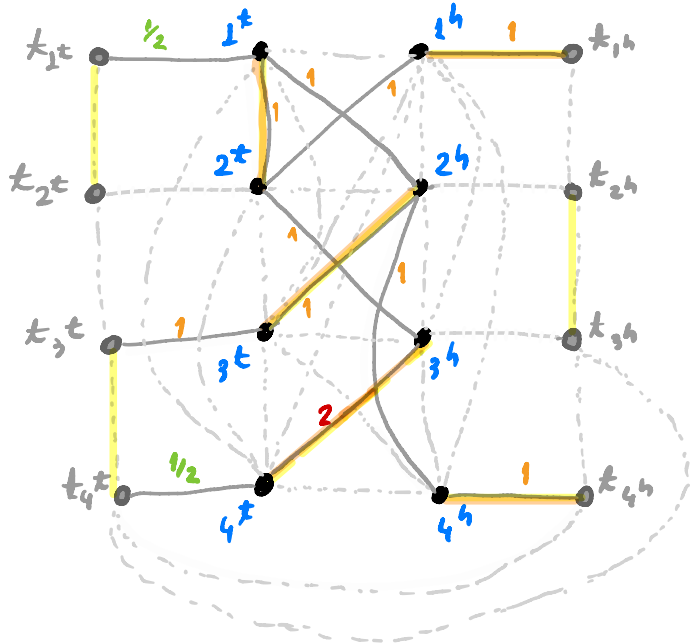
$$L_1 = [1 \ 2 \ 3 \ 4]$$

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

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

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

$$w(M) = 6$$



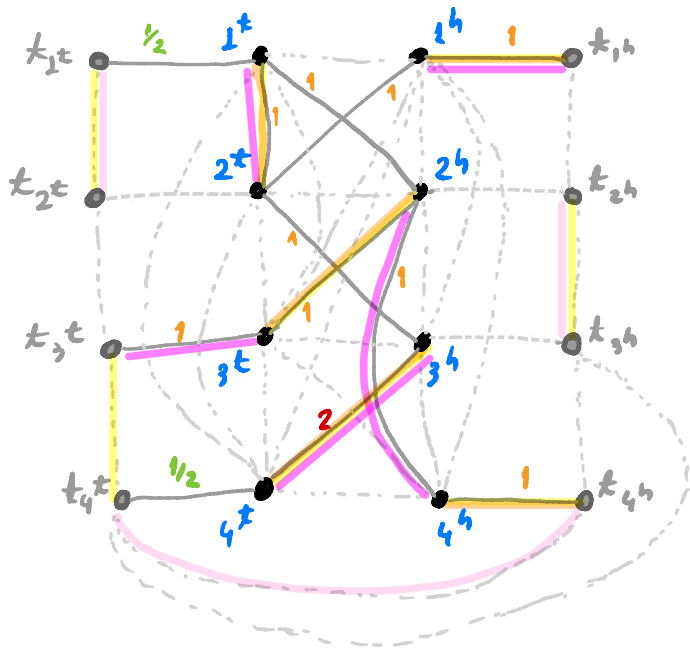
$$L_1 = [1 \ 2 \ 3 \ 4]$$

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

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

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

$$\omega(M_2) = 6$$



The breakpoint halving can
be computed in a similar way

Quiz 1

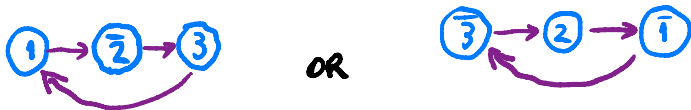
1 Which of the following statements are true?

- 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

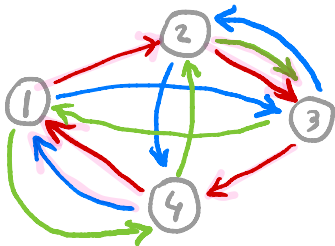
A unichromosomal circular genome \mathbb{C} can be represented as a simple directed cycle graph:

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



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: $\mathbb{C}_1^f = (1234)$, $\mathbb{C}_2^f = (2413)$, $\mathbb{C}_3^f = (2314)$



Every vertex has indegree =
outdegree = 3

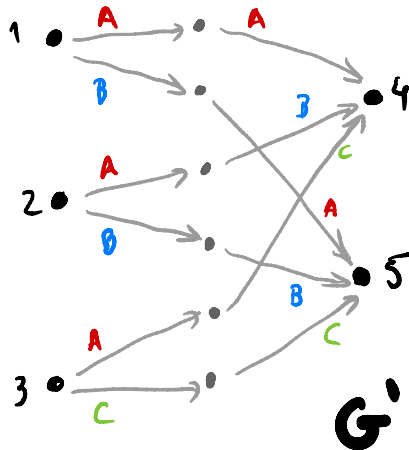
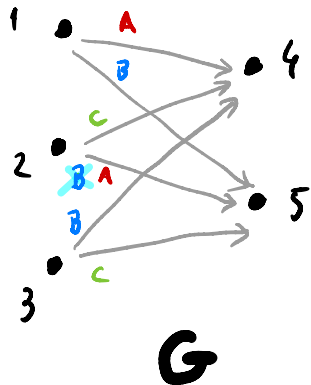
$$M = (1 \ 2 \ 3 \ 4)$$

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.

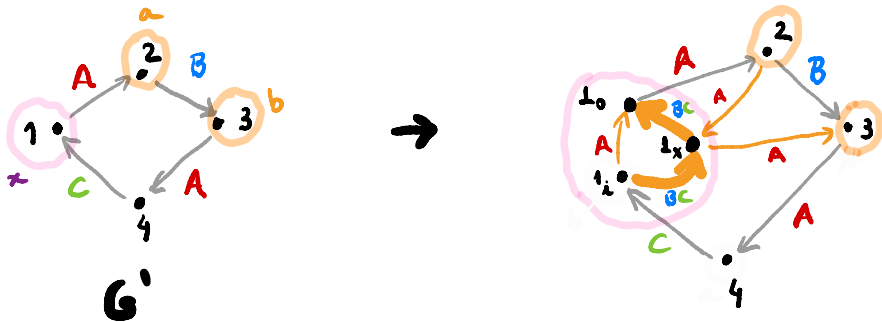
Reduction of this problem to the problem of computing a breakpoint median of three canonical circular genomes A , B and C that have the same relative orientation:

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 A, B and C)



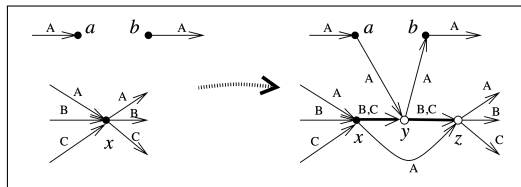
G'' has only adjacencies that occur in one or in two genomes

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

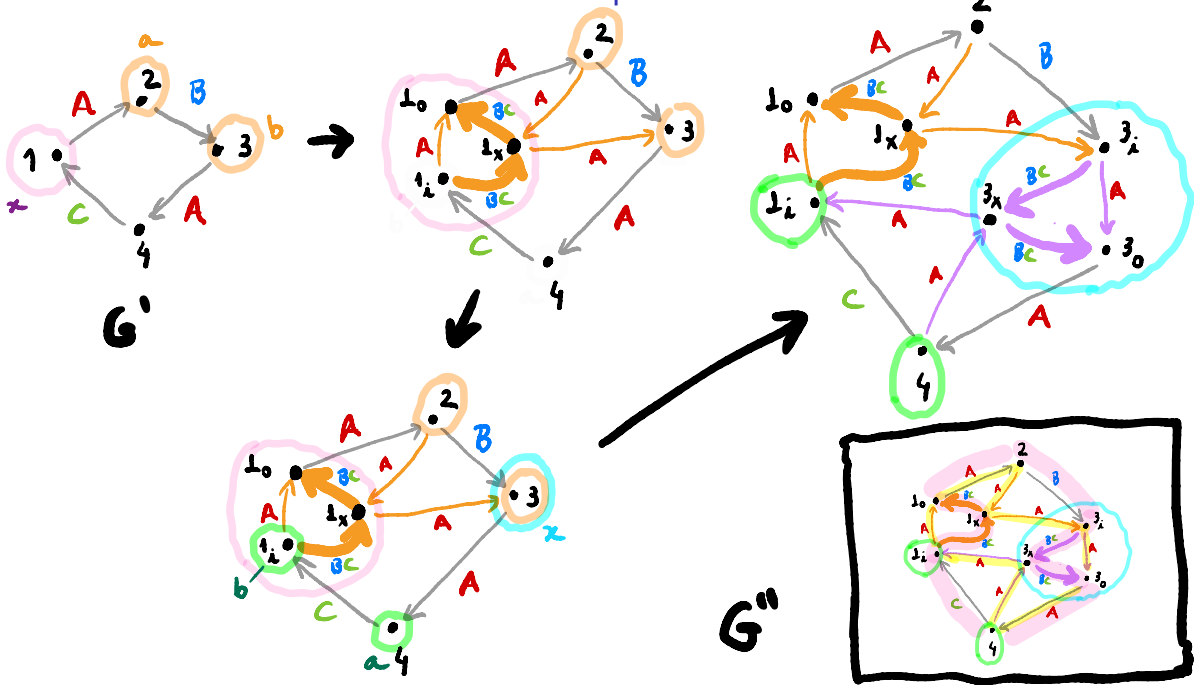
\mathbb{M} contains all adjacencies common to two input genomes and no "new" adjacency



Initial graph G has an hamiltonian cycle



NP-hardness of unichromosomal breakpoint median



Quiz 2

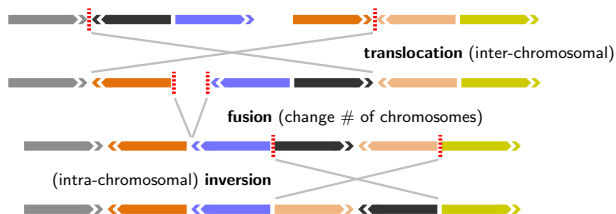
1 Which of the following statements are true?

- 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

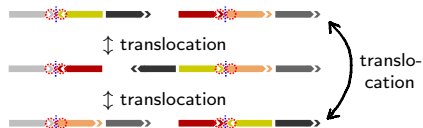
Double-cut-and-join (DCJ) operation: two cuts + two joins

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



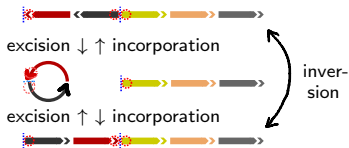
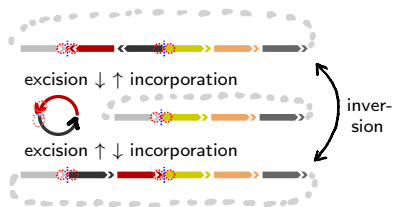
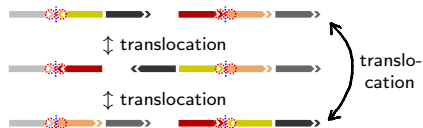
The double-cut-and-join (DCJ) operation

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



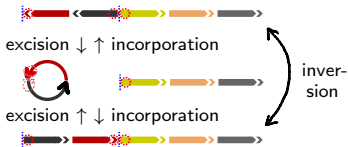
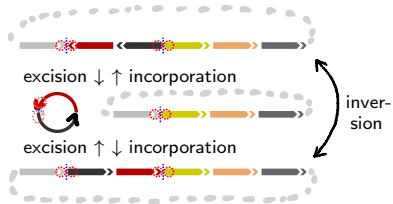
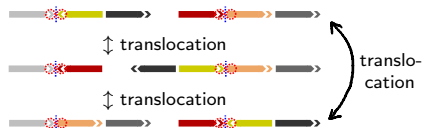
The double-cut-and-join (DCJ) operation

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



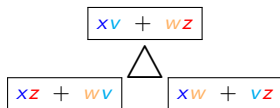
The double-cut-and-join (DCJ) operation

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



DCJ model

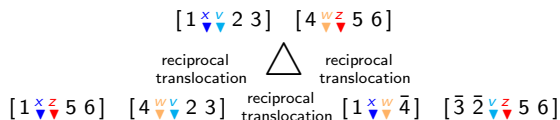
DCJ operation
involving
two adjacencies



two possibilities
of rejoining
in a different way

Cases:

A. Each adjacency is in a distinct linear chromosome:

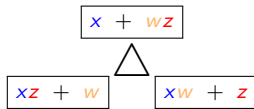


B. Both adjacencies are in the same chromosome, or one is in a circular chromosome:



DCJ model

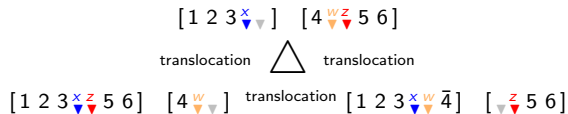
**DCJ operation
involving one adjacency
and one telomere**



**two possibilities
of rejoining
in a different way**

Cases:

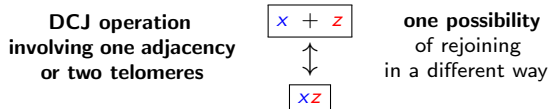
A. The adjacency and the telomere are in distinct linear chromosomes:



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

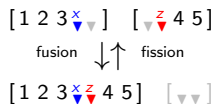


DCJ model



Cases:

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



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



Quiz 3

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

A $[123] [45] \leftrightarrow [12453]$

B $[123] [45] \leftrightarrow [123\bar{5}\bar{4}]$

C $[123] [45] \leftrightarrow [125] [43]$

D $[12345] \leftrightarrow [1\bar{4}3\bar{2}5]$

E $[12345] \leftrightarrow [12\bar{5}\bar{4}\bar{3}]$

F $[123] (45) \leftrightarrow [12453]$

G $[123] (45) \leftrightarrow [12543]$

H $(12345) \leftrightarrow [34512]$

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