

# Algorithms in Comparative Genomics

`gi.cebitec.uni-bielefeld.de/teaching/2021winter/cg`

## **Lecture:**

Marília D. V. Braga  
Thursdays, 10:15-11:45

## **Tutorial:**

Leonard Bohnenkämper  
Thursdays, 8:30-10:00

## Topics:

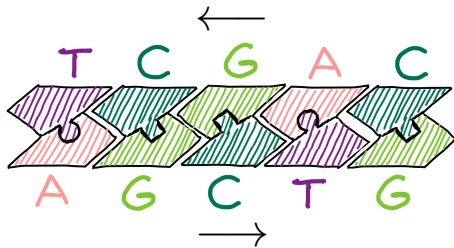
1. Genomes as gene orders or list of adjacencies
2. Family-annotated genomes, types of genomes
3. Large-scale rearrangements
4. Genome comparison problems: distance, double distance, median, halving
5. Breakpoint model / Single-cut-or-join (SCJ) model
6. Relational diagram of two genomes
7. Double-cut-and-join (DCJ) model
8. Inversion model
9. DCJ-indel model
10. NP-hard problems and ILP
  - 10.1 DCJ distance of balanced genomes
  - 10.2 DCJ-indel distance of natural genomes
  - 10.3 DCJ-indel distance of family-free genomes
11. Inferring gene families via family-free rearrangements
12. SCJ Small parsimony

## Topics of today - Introduction:

1. Genomes as gene orders or list of adjacencies
2. Family-annotated genomes, types of genomes
3. Large-scale rearrangements
4. Breakpoint distance, breakpoint double distance

## Each chromosome is a DNA molecule

The DNA molecule is a chain of oriented **base pairs** (bp)



Reverse complement:

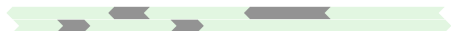
AGCTG  $\leftrightarrow$  CAGCT

(two complementary anti-parallel strands, linear or circular)

## Linear chromosomes as marker orders

Marker: oriented DNA fragment (lies on one of the two complementary anti-parallel DNA strands)

DNA breakpoints: between markers



↓ A chromosome is represented by its marker order



$$[ \quad \mathbb{A}[1] \quad \overline{\mathbb{A}[2]} \quad \mathbb{A}[3] \quad \overline{\mathbb{A}[4]} \quad ]$$

Set of markers:  $\mathcal{G}(\mathbb{A}) = \{ \mathbb{A}[1], \mathbb{A}[2], \mathbb{A}[3], \mathbb{A}[4] \}$

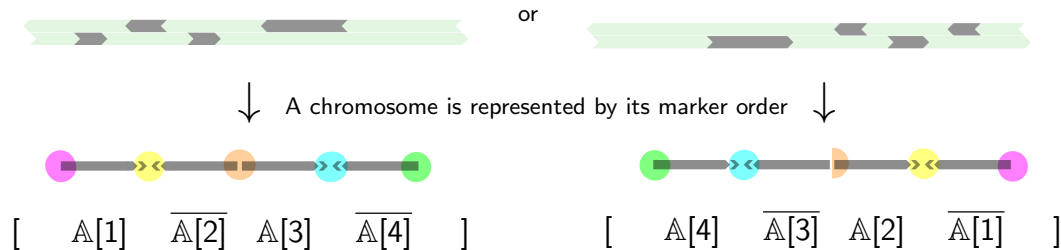
Set of adjacencies:  $\Gamma(\mathbb{A}) = \{ \mathbb{A}[1]^h \mathbb{A}[2]^h, \mathbb{A}[2]^t \mathbb{A}[3]^t, \mathbb{A}[3]^h \mathbb{A}[4]^h \}$

Set of telomeres:  $\Theta(\mathbb{A}) = \{ \mathbb{A}[1]^t, \mathbb{A}[4]^t \}$

## Linear chromosomes as marker orders

Marker: oriented DNA fragment (lies on one of the two complementary anti-parallel DNA strands)

DNA breakpoints: between markers

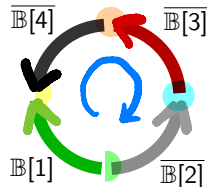
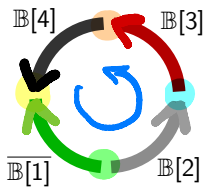


Set of markers:  $\mathcal{G}(A) = \{ A[1], A[2], A[3], A[4] \}$

Set of adjacencies:  $\Gamma(A) = \{ A[1]^h A[2]^h, A[2]^t A[3]^t, A[3]^h A[4]^h \}$

Set of telomeres:  $\Theta(A) = \{ A[1]^t, A[4]^t \}$

## Circular chromosomes as marker orders



$$( \overline{\mathbb{B}[1]} \quad \mathbb{B}[2] \quad \mathbb{B}[3] \quad \mathbb{B}[4] )$$

$$( \overline{\mathbb{B}[4]} \quad \overline{\mathbb{B}[3]} \quad \overline{\mathbb{B}[2]} \quad \mathbb{B}[1] )$$

$$( \mathbb{B}[4] \quad \overline{\mathbb{B}[1]} \quad \mathbb{B}[2] \quad \mathbb{B}[3] )$$

$$( \overline{\mathbb{B}[3]} \quad \overline{\mathbb{B}[2]} \quad \mathbb{B}[1] \quad \overline{\mathbb{B}[4]} )$$

$$( \mathbb{B}[3] \quad \mathbb{B}[4] \quad \overline{\mathbb{B}[1]} \quad \mathbb{B}[2] )$$

$$( \overline{\mathbb{B}[2]} \quad \mathbb{B}[1] \quad \overline{\mathbb{B}[4]} \quad \overline{\mathbb{B}[3]} )$$

$$( \mathbb{B}[2] \quad \mathbb{B}[3] \quad \mathbb{B}[4] \quad \overline{\mathbb{B}[1]} )$$

$$( \mathbb{B}[1] \quad \overline{\mathbb{B}[4]} \quad \overline{\mathbb{B}[3]} \quad \overline{\mathbb{B}[2]} )$$

Set of markers:  $\mathcal{G}(\mathbb{B}) = \{ \mathbb{B}[1], \mathbb{B}[2], \mathbb{B}[3], \mathbb{B}[4] \}$

Set of adjacencies:  $\Gamma(\mathbb{B}) = \{ \mathbb{B}[1]^t \mathbb{B}[2]^t, \mathbb{B}[2]^h \mathbb{B}[3]^t, \mathbb{B}[3]^h \mathbb{B}[4]^t, \mathbb{B}[4]^h \mathbb{B}[1]^h \}$

## Family annotated genome



Set of families:  $\mathcal{F}(\mathbb{A}^f) = \{ 1, 2, 3, 4, 5 \}$

Multiset of genes:  $\mathcal{G}(\mathbb{A}^f) = \{ 1, 1, 2, 3, 4, 4, 5 \}$

Multiset of adjacencies:  $\Gamma(\mathbb{A}^f) = \{ 1^h 2^h, 2^t 3^t, 4^h 1^h, 1^t 4^t, 4^h 5^h \}$

Multiset of telomeres:  $\Theta(\mathbb{A}^f) = \{ 1^t, 3^h, 4^t, 5^t \}$



# Types of genomes

- ▶ Unichromosomal × multichromosomal
- ▶ Linear, circular, mixed
- ▶ Concerning the gene content:

1. **Singular genome**  $G_{\triangleright}^f$ : each family occurs **exactly once**



2. **Duplicated genome**  $G_{\circ}^f$ : each family occurs **exactly twice**



3. **Perfectly duplicated or doubled genome**  $G_{\boxtimes}^f$ : each adjacency or telomere occurs **exactly twice**



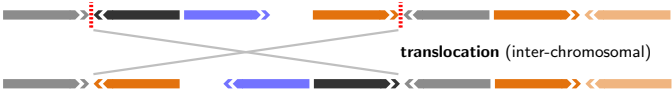
4. **Natural genome**: **no restriction** on the number of occurrences of families



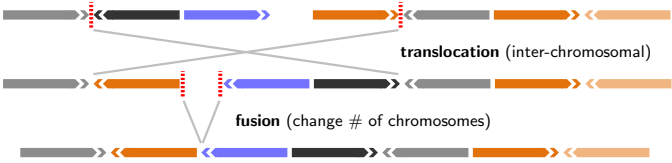
## Large-scale genome rearrangements



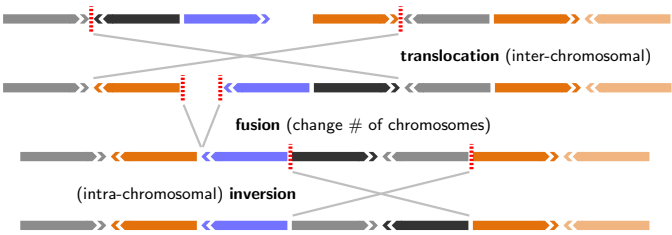
# Large-scale genome rearrangements



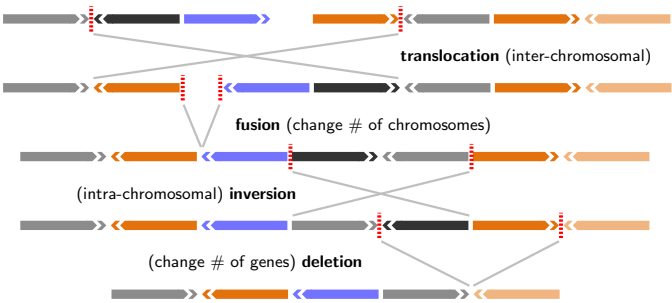
# Large-scale genome rearrangements



# Large-scale genome rearrangements



# Large-scale genome rearrangements



# Comparison of genomes



↕ compute a distance measure  
between two genomes

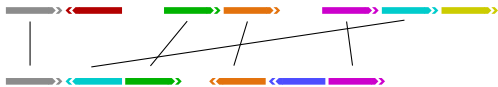


$$d(A^f, B^f)$$

# Types of genome pairs

## Pair of singular genomes:

each family occurs **at most once** in each genome



## Pair of balanced genomes:

each family occurs **the same number of times** in each genome

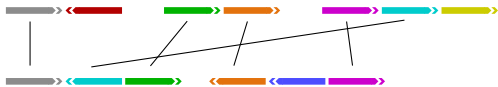




# Types of genome pairs

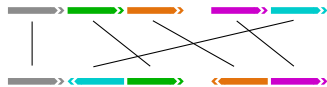
## Pair of singular genomes:

each family occurs **at most once** in each genome



## Pair of canonical genomes:

singular and balanced



## Pair of balanced genomes:

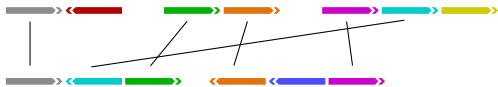
each family occurs **the same number of times** in each genome



# Types of genome pairs

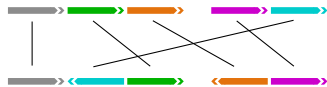
## Pair of singular genomes:

each family occurs **at most once** in each genome



## Pair of canonical genomes:

singular and balanced



## Pair of balanced genomes:

each family occurs **the same number of times** in each genome



## Pair of natural genomes:

**no restriction** on the number of occurrences of families

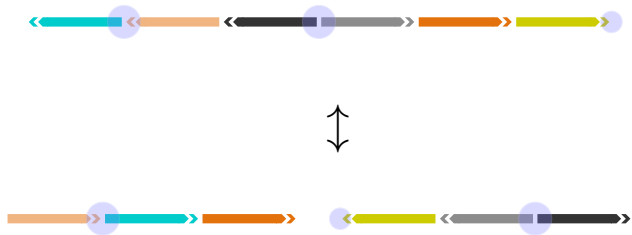


Resolving ambiguous families with a maximal matching

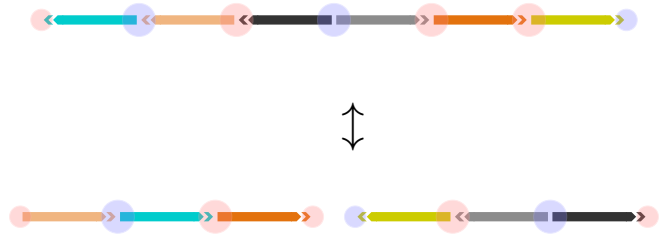
# Canonical genomes: common adjacency $\times$ breakpoint



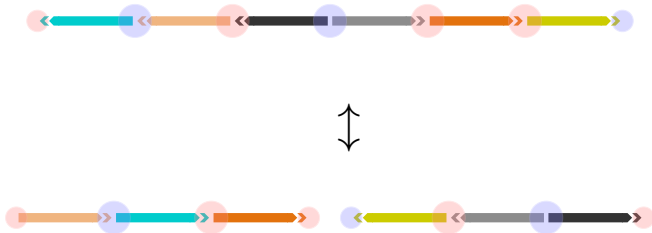
# Canonical genomes: common adjacency $\times$ breakpoint



Canonical genomes: common adjacency  $\times$  breakpoint



# Breakpoint distance of canonical genomes



$$d_{BP}(A_{\triangleright}^f, B_{\triangleright}^f) = n - a - \frac{t}{2}$$

## Obtaining doubled genomes from a singular genome

Given a singular genome  $\mathbb{G}_{\triangleright}^f$ , let  $2 \cdot \mathbb{G}_{\triangleright}^f$  be the set of doubled genomes obtained by duplicating each adjacency and each telomere of  $\mathbb{G}_{\triangleright}^f$ .

Examples:

$$\mathbb{A}_{\triangleright}^f = \{ [\bar{2}13] \} \quad \Rightarrow \quad 2 \cdot \mathbb{A}_{\triangleright}^f = \{ \{ [\bar{2}13] [\bar{2}13] \} \}$$

$$\mathbb{B}_{\triangleright}^f = \{ (\bar{2}13) \} \quad \Rightarrow \quad 2 \cdot \mathbb{B}_{\triangleright}^f = \{ \{ (\bar{2}13) (\bar{2}13) \}, \{ (\bar{2}13\bar{2}13) \} \}$$

$$\mathbb{C}_{\triangleright}^f = \{ (2) (\bar{1}3) \} \quad \Rightarrow \quad 2 \cdot \mathbb{C}_{\triangleright}^f = \{ \{ (2) (2) (\bar{1}3) (\bar{1}3) \}, \{ (22) (\bar{1}3) (\bar{1}3) \}, \\ \{ (2) (2) (\bar{1}3\bar{1}3) \}, \{ (22) (\bar{1}3\bar{1}3) \} \}$$



## Breakpoint double distance

Given a singular genome  $\mathbb{A}_{\triangleright}^f$  and a duplicated genome  $\mathbb{B}_{\diamond}^f$ ,  
the **breakpoint double distance** is defined as:

$$d_{\text{BP}}^2(\mathbb{A}_{\triangleright}^f, \mathbb{B}_{\diamond}^f) = \min_{\mathbb{A}_{\triangleright\triangleleft}^f \in 2 \cdot \mathbb{A}_{\triangleright}^f} d_{\text{BP}}(\mathbb{A}_{\triangleright\triangleleft}^f, \mathbb{B}_{\diamond}^f)$$

Ex:  $\mathbb{A}_{\triangleright}^f = [\bar{2} \ 1 \ \bar{3}]$  and  $\mathbb{B}_{\diamond}^f = [3\bar{1} \ \bar{2} \ 3 \ \bar{1} \ 2]$



# Quiz

Given genomes  $\mathbb{A}^f = (1234) [1\bar{5}\bar{4}5\bar{3}\bar{2}]$ ,  $\mathbb{B}_{\triangleright}^f = [12345]$  and  $\mathbb{C}_{\triangleright}^f = [\bar{2}\bar{1}] [\bar{4}\bar{3}5]$ .

1 Which of the following statements are true?

A Genome  $\mathbb{A}$  is linear.

B Genome  $\mathbb{A}$  is multichromosomal.

C Genome  $\mathbb{A}^f$  is duplicated.

D Genome  $\mathbb{A}^f$  is doubled.

2 How many families occur in genome  $\mathbb{A}^f$ ?

A 4

B 5

C 5.5

D 6

3 What is the breakpoint distance of  $\mathbb{B}_{\triangleright}^f$  and  $\mathbb{C}_{\triangleright}^f$ ?

A 1.5

B 2

C 2.5

D 3

4 What is the breakpoint double distance of  $\mathbb{A}^f$  and  $\mathbb{B}_{\triangleright}^f$ ?

A 4

B 4.2

C 4.5

D 5

## Reference

Multichromosomal median and halving problems under different genomic distances

(Eric Tannier, Chunfang Zheng and David Sankoff)

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