Algorithms in Comparative Genomics

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

Tutorial:
Leonard Bohnenkämper
Thursdays (or Fridays?), 8:30-10:00

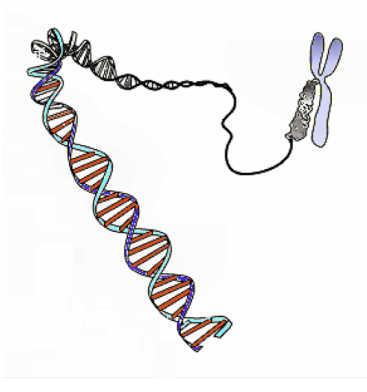
Topics:

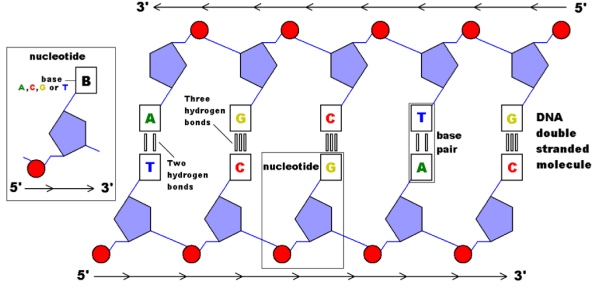
1. Genomes as gene orders or list of adjacencies
2. Large-scale rearrangements
3. Types of genomes
4. Breakpoint distance, breakpoint double distance
5. Common intervals and gene clusters
6. Relational diagram of two genomes
7. Double-cut-and-join (DCJ) operation
   - 7.1 DCJ distance
   - 7.2 DCJ-indel distance
   - 7.3 Capping (circularizing genomes)
   - 7.4 Duplications and ILP
   - 7.5 Family-free setting
8. Inversion distance
9. Single-cut-or-join (SCJ) distance and median
10. Small parsimony

Topics of today - Introduction:

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

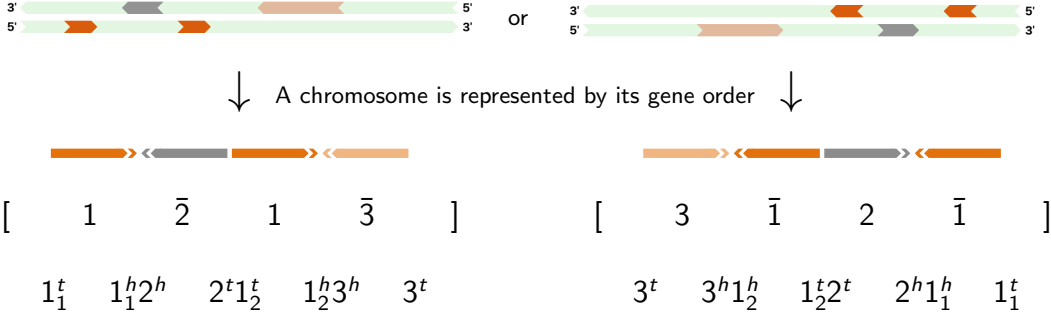
Genomes




3'
5'
nucleotide
base
A, C, G or T
B
5'
3'
Three hydrogen bonds
A
G
C
T
base pair
G
DNA double stranded molecule
T
Two hydrogen bonds
C
nucleotide
G
A
C
5'
3'

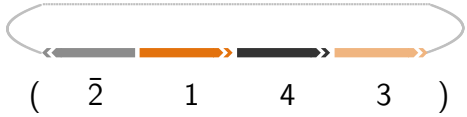
Genomes as gene orders or list of adjacencies - linear chromosomes

Each gene is an oriented DNA fragment:
it lies on one of the two complementary anti-parallel DNA strands



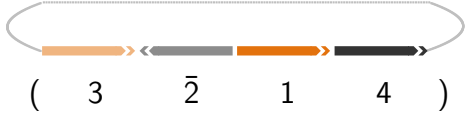
(genes of the same color/number belong to the same **family**)

Genomes as gene orders or list of adjacencies - circular chromosomes

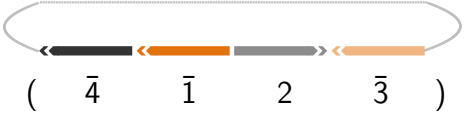


$(\ \bar{2}\ \ 1\ \ 4\ \ 3\ )$

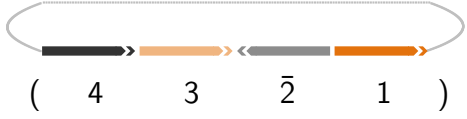
$(\ \bar{3}\ \ \bar{4}\ \ \bar{1}\ \ 2\ )$



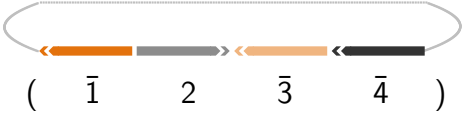
$(\ 3\ \ \bar{2}\ \ 1\ \ 4\ )$



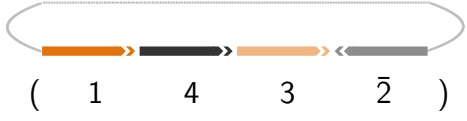
$(\ \bar{4}\ \ \bar{1}\ \ 2\ \ \bar{3}\ )$



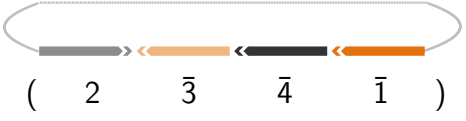
$(\ 4\ \ 3\ \ \bar{2}\ \ 1\ )$



$(\ \bar{1}\ \ 2\ \ \bar{3}\ \ \bar{4}\ )$



$(\ 1\ \ 4\ \ 3\ \ \bar{2}\ )$

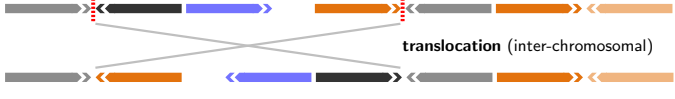


$(\ 2\ \ \bar{3}\ \ \bar{4}\ \ \bar{1}\ )$

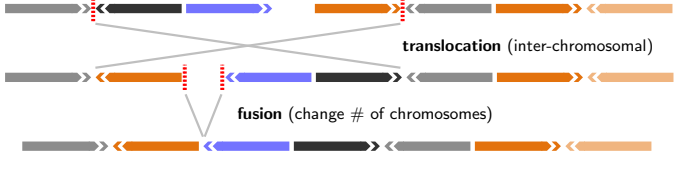
$1^h4^t \quad 4^h3^t \quad 3^h2^h \quad 2^t1^t$

Large-scale genome rearrangements

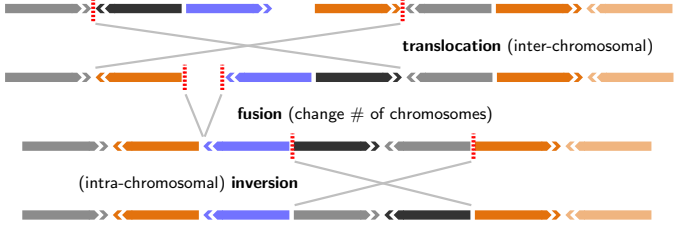
Large-scale genome rearrangements



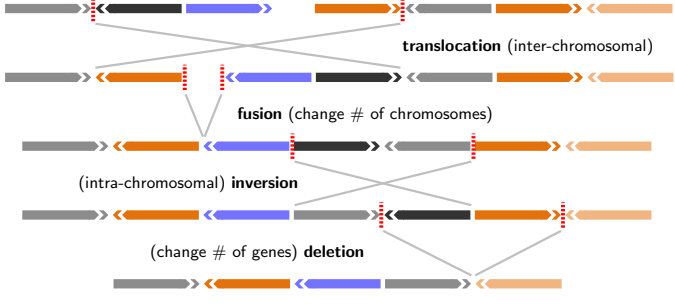
Large-scale genome rearrangements



Large-scale genome rearrangements



Large-scale genome rearrangements



Types of genomes

- Unichromosomal $\times$ multichromosomal
- Linear, circular, mixed
- Concerning the gene content:

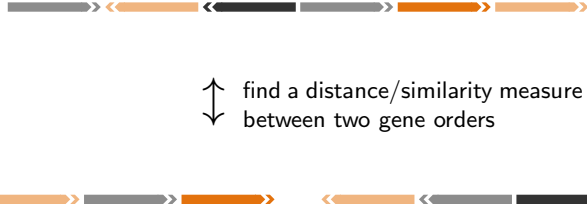
1. **Singular genome:** each family occurs **exactly once**

2. **Duplicated genome:** each family occurs **exactly twice**

3. **Perfectly duplicated or doubled genome:** each adjacency or telomere occurs **exactly twice**

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

Comparison of genomes



Types of genome pairs

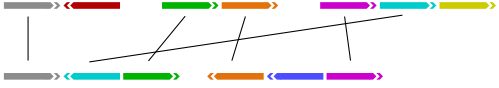
Pair of singular genomes:
each family occurs **at most once** in each genome

Pair of balanced genomes:
each family occurs **the same number of times** in each genome

Pair of natural genomes:
no restriction on the number of occurrences of families

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

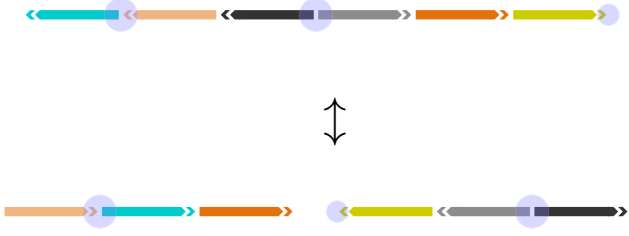
Pair of natural genomes:
no restriction on the number of occurrences of families

Breakpoint distance of canonical genomes

Common adjacency $\times$ breakpoint

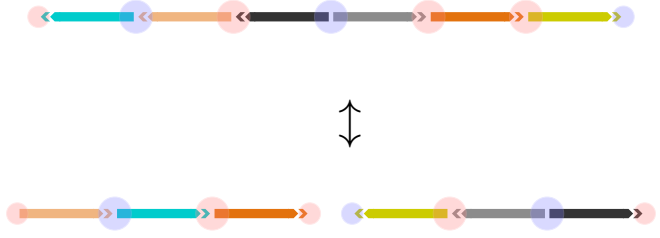
Breakpoint distance of canonical genomes

Common adjacency $\times$ breakpoint



Breakpoint distance of canonical genomes

Common adjacency $\times$ breakpoint



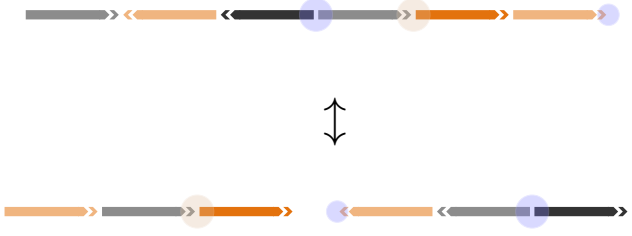
$$d_{\mathrm{BP}}(A, B) = n - a - \tfrac{t}{2}$$

Breakpoint distance of balanced genomes

Common adjacency $\times$ breakpoint

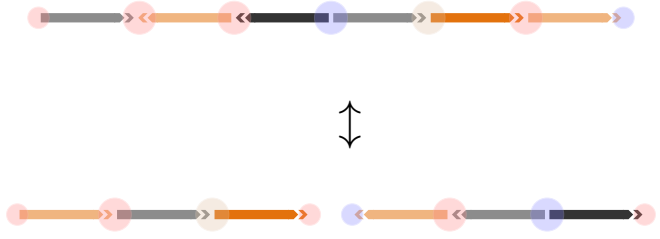
Breakpoint distance of balanced genomes

Common adjacency $\times$ breakpoint



Breakpoint distance of balanced genomes

Common adjacency $\times$ breakpoint



$$d_{\mathrm{BP}}(A, B) = n - a - \tfrac{t}{2}$$

Breakpoint double distance

Given a singular genome S, let $S \oplus S$ be a doubled genome
(obtained by duplicating each adjacency and each telomere of S).

Examples:

$S_1 = [\bar{2}\,1\,3]$ and $S_1 \oplus S_1 = [\bar{2}\,1\,3] \;\; [\bar{2}\,1\,3]$

$S_2 = (\bar{2}\,1\,3)$ and $S_2 \oplus S_2 = (\bar{2}\,1\,3) \;\; (\bar{2}\,1\,3)$ or $S_2 \oplus S_2 = (\bar{2}\,1\,3\,\bar{2}\,1\,3)$

Given a duplicated genome D, the breakpoint double distance is defined as:

$$d_{\mathrm{BP}}^2(D, S) = d_{\mathrm{BP}}(D, S \oplus S)$$

Ex: $D = [\bar{1}\,2\,\bar{3}\,1\,3\,2]$ and $S = [\bar{2}\,1\,3]$

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)

Quiz

Given genomes $A = (1\,2\,3\,4)\ [1\,\bar{5}\,\bar{4}\,5\,\bar{3}\,\bar{2}]$, $B = [1\,2\,3\,4\,5]$ and $C = [\bar{2}\,\bar{1}]\ [\bar{4}\,\bar{3}\,5]$.

1 Which of the following statements are true?

A Genome A is linear.

B Genome A is multichromosomal.

C Genome A is duplicated.

D Genome A is doubled.

2 What is the breakpoint distance of B and C?

A 1,5

B 2

C 2,5

D 3

2 How many families occur in genome A?

A 4

B 5

C 5,5

D 6

4 What is the breakpoint double distance of A and B?

A 4

B 4,2

C 4,5

D 5

Breakpoint halving

Given a duplicated genome D,
find a singular genome S that minimizes
the breakpoint double distance:

$$d_{\mathrm{BP}}^2(D, S) = d_{\mathrm{BP}}(D, S \oplus S)$$