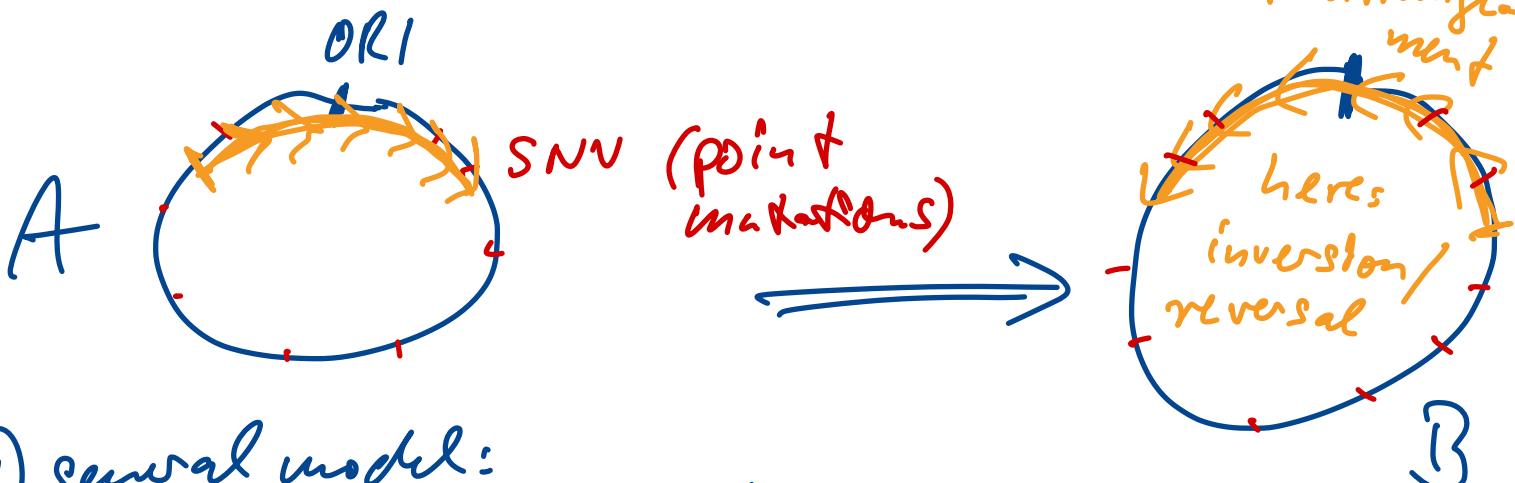


Jan. 29, 2021

# Comparative Genomics I

## - Genome Rearrangement



① general model:

- genes are represented by numbers from the set  $N = \{1, \dots, n\}$

- often genes are signed (+ / -)

→ a genome is a (signed) sequence over  $N$ .

Set of...

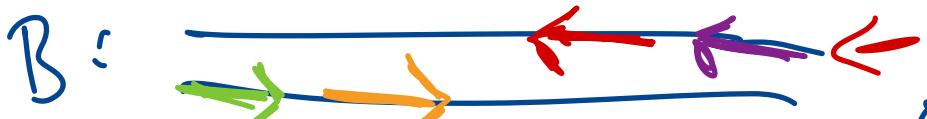
for multiple chromosomes

gene orders:



+1    -2    -3    +4 +1 or  $[1, -2, -3, 4]$

↑ comparison



+3    +2    -1    -4    -1

## ② different situations:

1) canonical genomes: each gene appears exactly once (like above)

2) singular genomes: each gene appears at most once

3) balanced genomes: each gene appears the same number of times in each studied genome

4) natural genomes: each gene may appear any number of times (or not at all)

## ③ Typical questions:

- evolutionary distance
- reconstruct phylogenies
- find regions of co-occurring genes (gene clusters)
- preprocessing of genome alignment

# Benzene Rearrangement Distances

## ① evolution w/s operations

- inversion / reversal



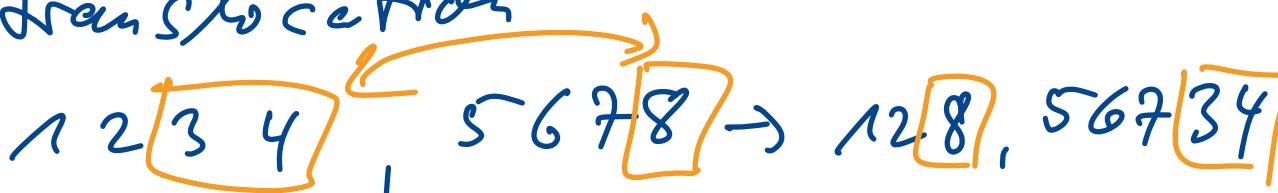
- transposition



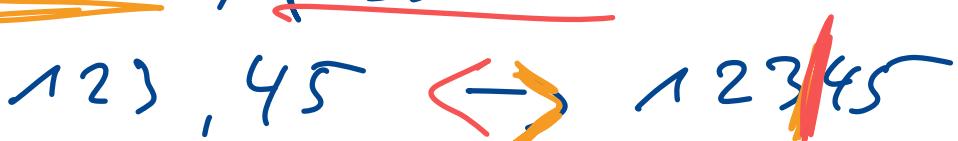
- block interchange



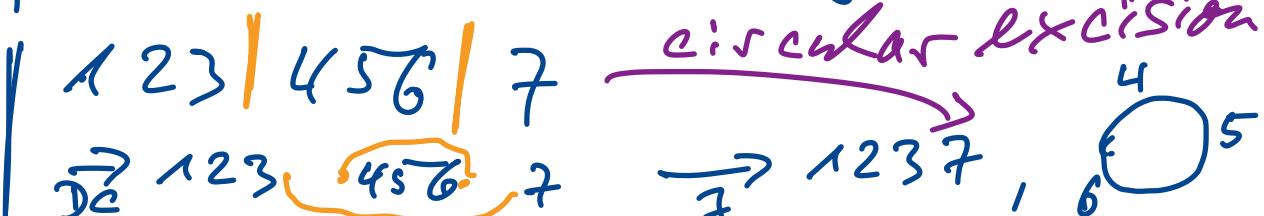
- translocation

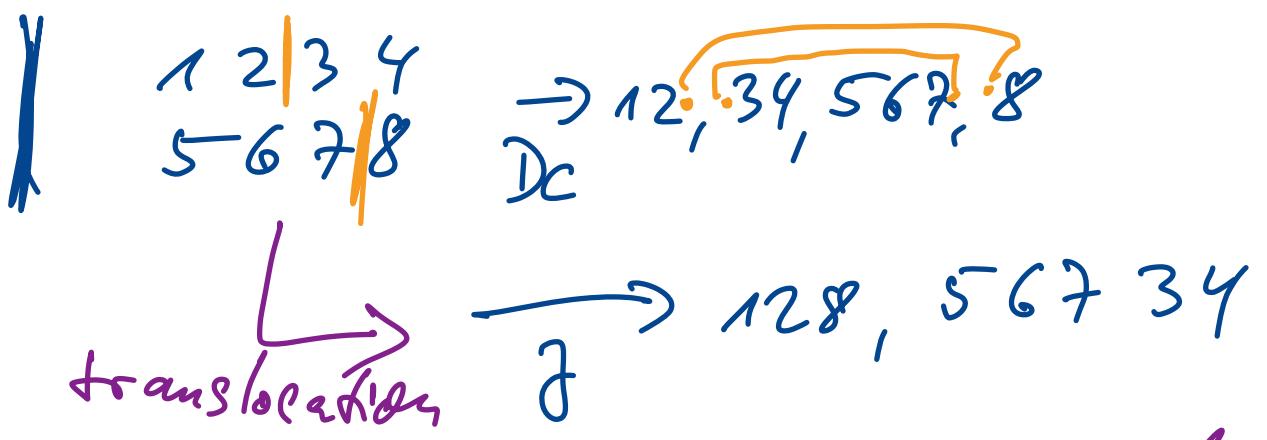


- fusion, fission



- double-cut and join (DCJ)





also: fusion, fission, circular insertion

- single cut or join (SCJ)

$$123|4 \rightarrow 123, 4$$

$$12\cdot 34 \rightarrow 12;34$$

## ② genomic distances: (Transformation distances)

- SCJ distance:  $d_{scj}(A, B)$  is the minimum number of SCJ operations necessary to transform genome A into genome B.

e.g.  $A = [1, 2], [3 | 4 | 5]$   
 $B = [1, 2, 3] [4] [5]$

$$d_{scj}(A, B) = 3$$

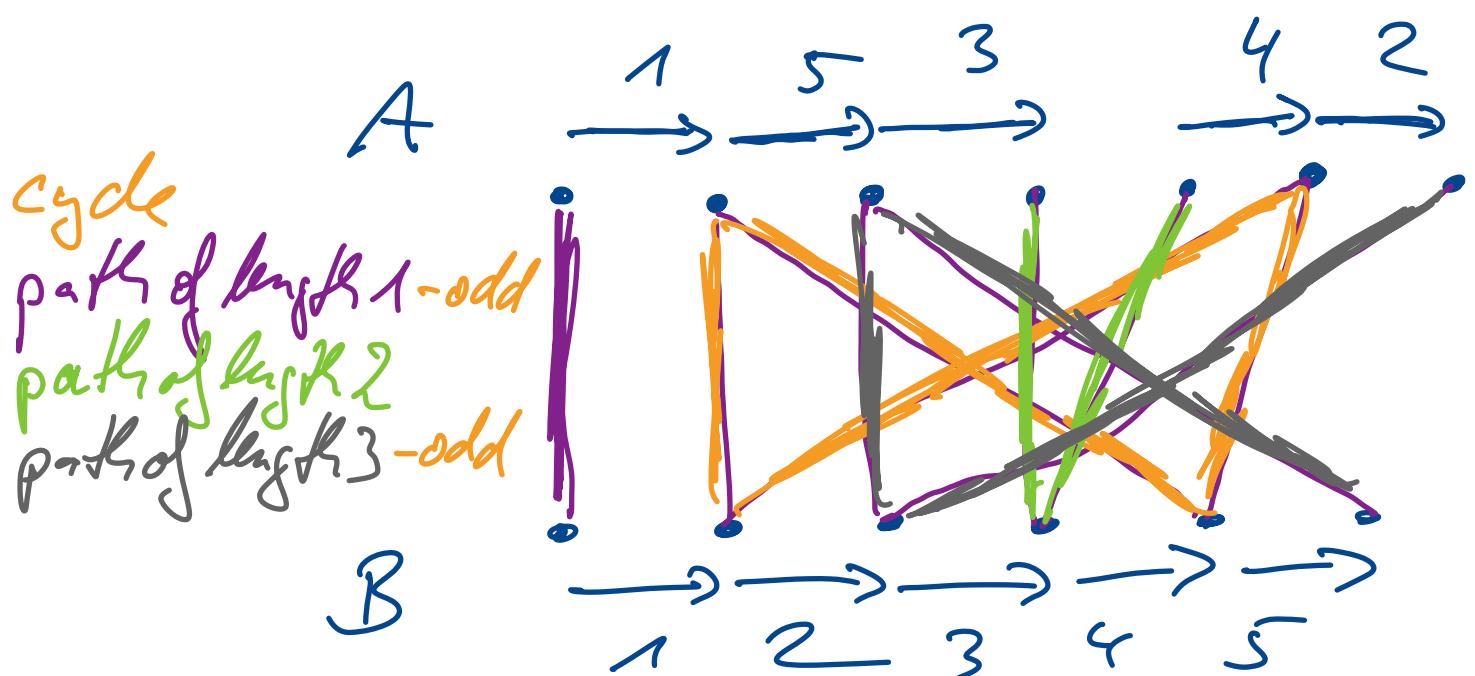
- DCJ distance  $d_{dcj}(A, B)$  is ..

- reversal (inversion) distance  $d_R(A, B)$ ...
- transposition distance  $d_T(A, B)$ ...
- transposition + reversal distance  $d_{TR}(A, B)$
- translocation distance  $d_H(A, B)$
- translocation + reversal distance  
 $d_{HP}(A, B)$

Hannenhalli & Pevzner  
1995

## DCJ distance for canonical genomes

def: adjacency graph:  
 1 vertex for each adjacencies  
 + 1 vertex for each telomere



Theorem:

$$d_{DCJ}(A, B) = n - c - \frac{i}{2} = 5 - 1 - \frac{1}{2} = 3$$

where  $n = \# \text{ genes}$  (here: 5)

$c = \# \text{ cycles in the adjacency graph} = 1$

$i = \# \text{ odd paths in the adj.: graph} = 2$

Algorithm:  $O(n)$  time & space

③ static distances

$d_{BP}(AB)$

$$A: [x \cdot 1 \cdot 5 \cdot 6 \cdot 3], [4 \cdot 2 \cdot x]$$

$$\underline{d_{BP}(A,B) = 4\frac{1}{2}}$$

$$B: [x \cdot 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot x]$$

$4\frac{1}{2}$

$\parallel$

$4\frac{1}{2}$

breakpoint distance

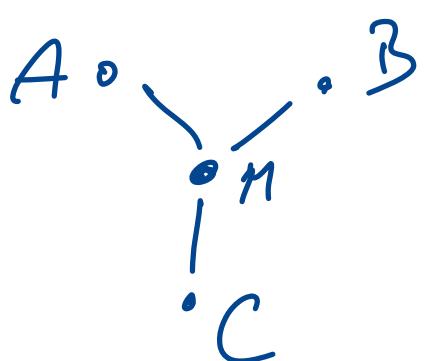
$\begin{array}{l} \text{adj}^+ = +1 \\ \text{tel}^- = +1\frac{1}{2} \end{array}$

$\neq$  conserved adjacency

+0

beyond distances :

median of three geodesics



given  $A, B, C,$   
find  $M$   
such that  
 $d(A,M) + d(M,B) + d(B,C)$   
 $\rightarrow \min.$

results:

1) for most distances  
(e.g.  $d_{DCJ}$ ,  $d_{REV}$ ,  $d_{AP}$ , ...)  
the median problem is NP-hard.

2) for SCJ the median problem  
is polynomial.

method: count adjacencies in  $A, B, C$ :

if any adjacency is conserved  
2 or 3 times, then add it to  $M$ ;  
otherwise not.