CENTER GENOME WITH RESPECT TO THE RANK DISTANCE

Priscila Biller João P. P. Zanetti João Meidanis

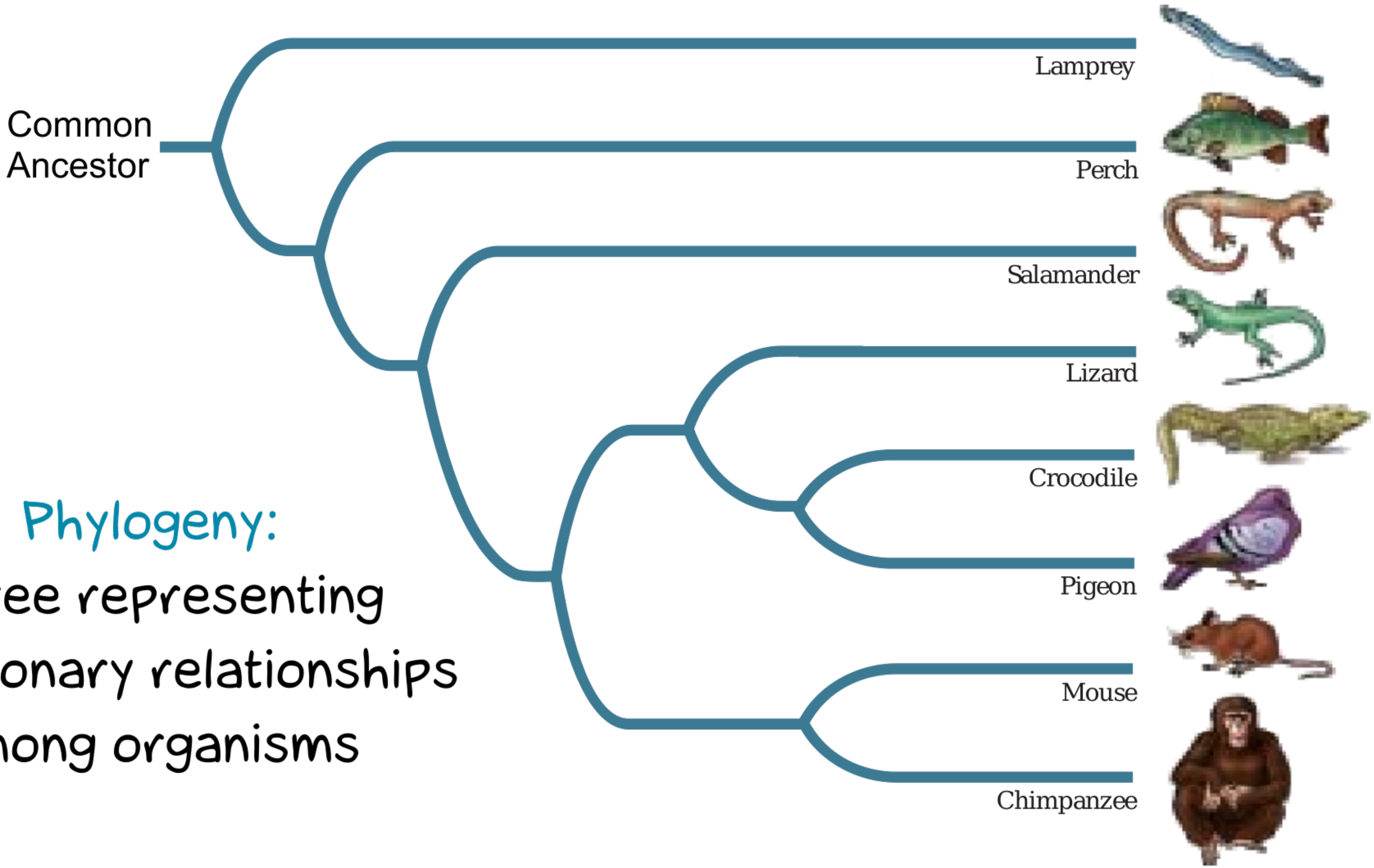
UNICAMP

Scylla INFORMATICS

FAPESP

SFU SIMON FRASER UNIVERSITY

How to infer ancestral genomes?



Phylogeny:

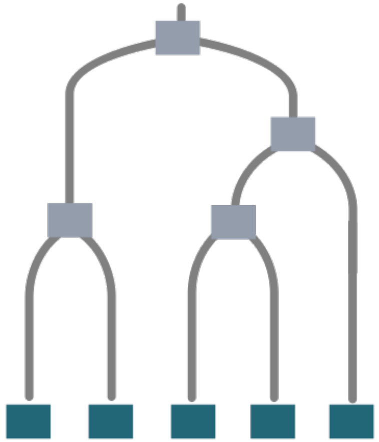
a tree representing evolutionary relationships among organisms

Internal nodes:

ancestors (usually extinct)

Leaves:

recent species (known genomes)



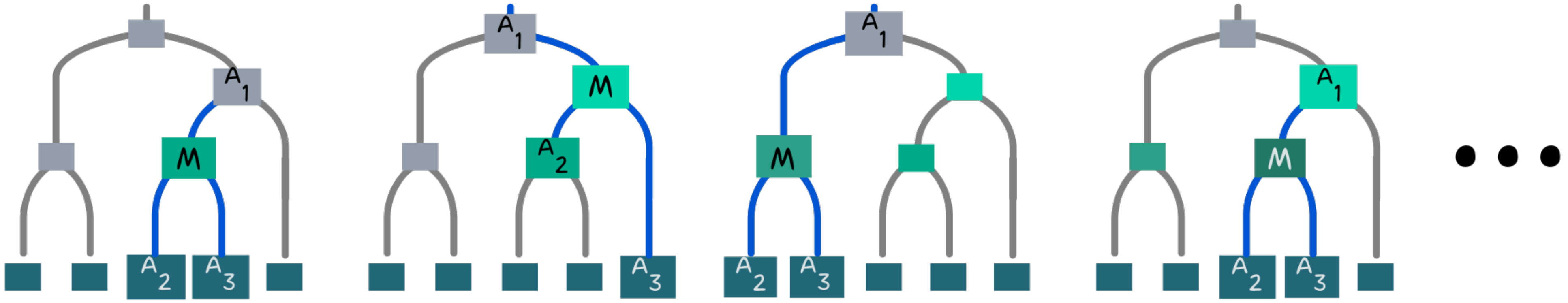
How to infer ancestral genomes?

■ Known genomes

■ Genomes to be inferred (initially arbitrary)

Usual way to infer ancestors:

Repeatedly compute the median genome M until convergence is reached



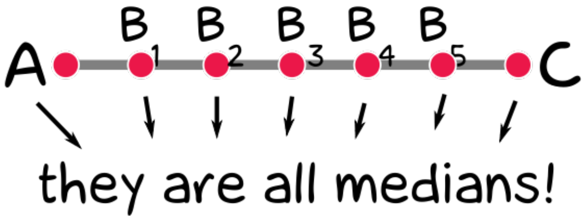
Genome M minimizes the sum of the evolutionary distances:

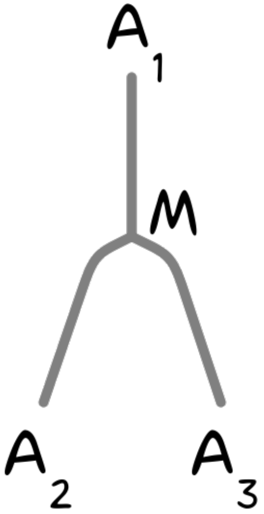
$$d(A_1, M) + d(A_2, M) + d(A_3, M)$$

Are median genomes a good way to infer ancestors?

2 input genomes: solutions are not relevant

Any genome in an optimal sorting scenario minimizes d(A, B) + d(B, C), including the input genomes





3 or more input genomes: hard

NP-hard for most rearrangement distances (reversal, DCJ, etc.)

Center genome: an alternative to the median

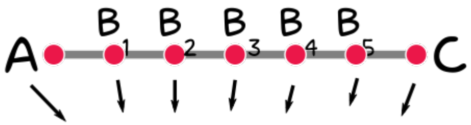
Median genome

Input: genomes $A_1, A_2, \dots$

Goal: find a genome M that minimizes

$d(A_1, M) + d(A_2, M) + d(A_3, M) + \dots$

2 input genomes:



they are all medians!

✗

3 or more input genomes: NP-hard

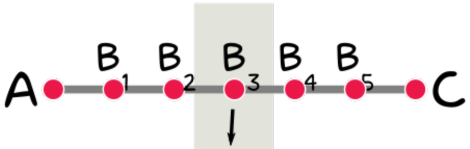
Center genome

Input: genomes $A_1, A_2, \dots$

Goal: find a genome M that minimizes

$\max(d(A_1, M), d(A_2, M), d(A_3, M), \dots)$

2 input genomes:



more constrained set of solutions

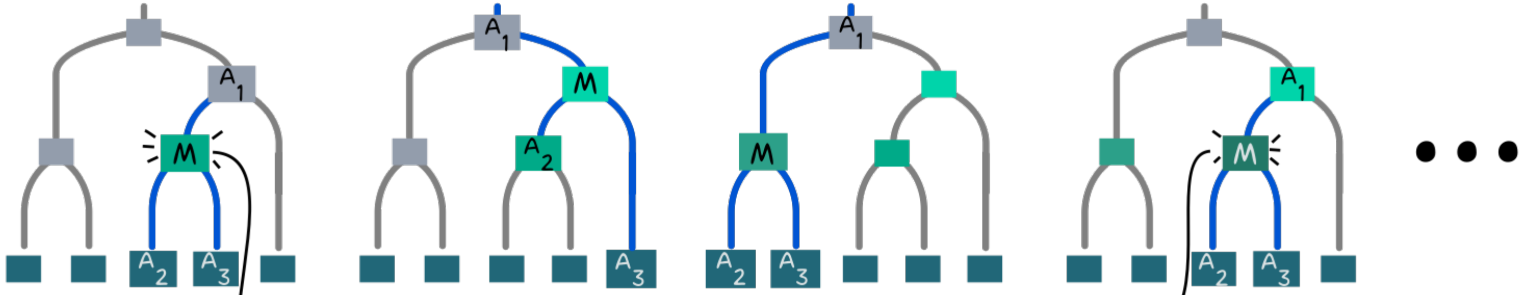
✓

3 or more input genomes: Open (NP-hard?)

Ancestral inference: center genomes are an appealing alternative to the median

Median of 3 genomes

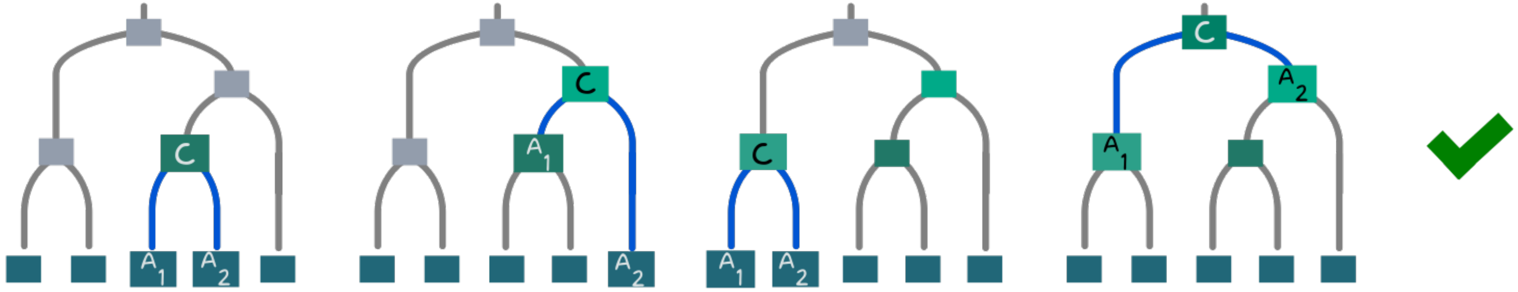
Hard



Recomputes the same ancestor several times until convergence is achieved

Center of 2 genomes

Easy



Direct method

Center genomes with respect to the rank distance

Gabidulin Ernst

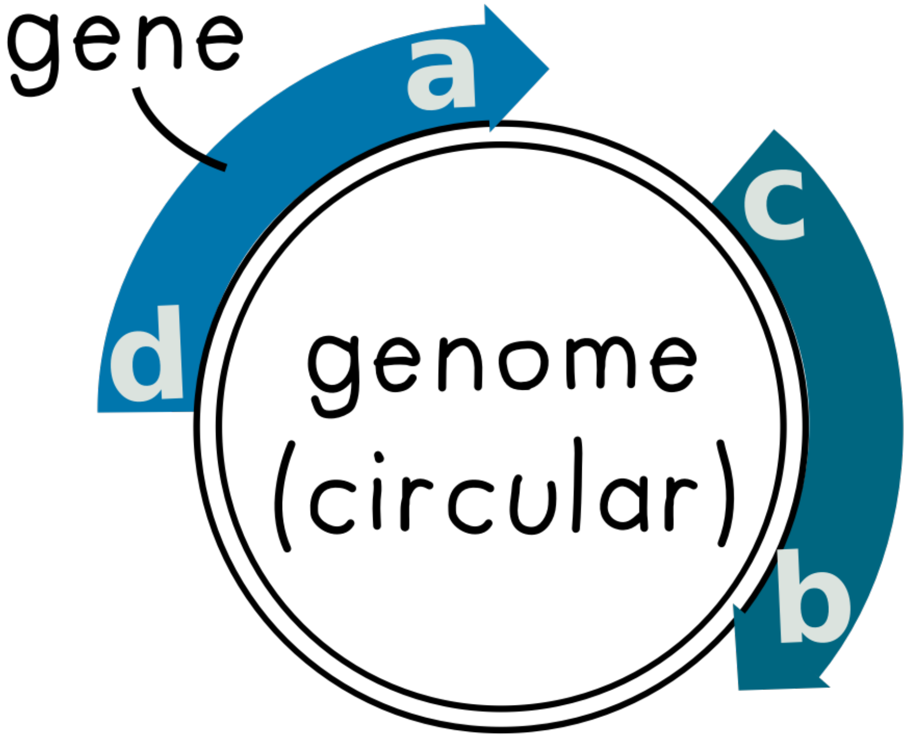
Rank distance:

very successfully used in coding theory since at least 1985

$$\boxed{d(\underbrace{A, B}_{\text{matrices}}) = \text{rank}(A-B)}$$

rank(X): dimension of the image (or column space) of X

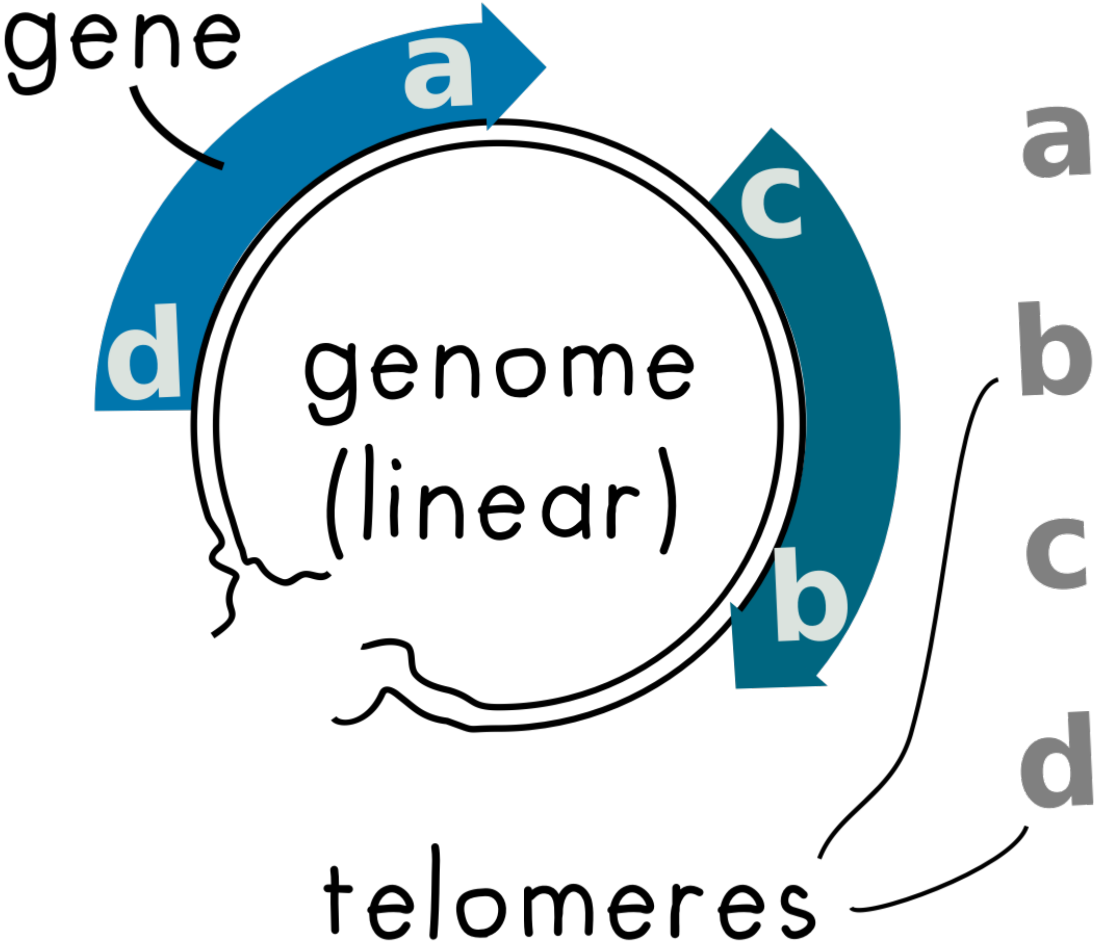
How to represent genomes as matrices?



$$
\begin{array}{c|cccc}
 & a & b & c & d \\
\hline
a & 0 & 0 & 1 & 0 \\
b & 0 & 0 & 0 & 1 \\
c & 1 & 0 & 0 & 0 \\
d & 0 & 1 & 0 & 0
\end{array}
$$

gene extremities

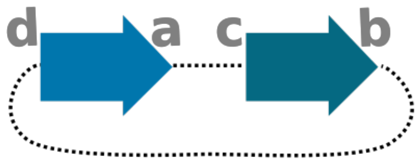
How to represent genomes as matrices?



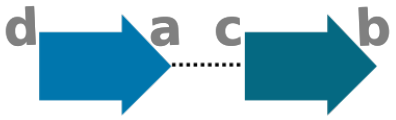
$$\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 0 & 1 & 0 \\ b & 0 & 1 & 0 & 0 \\ c & 1 & 0 & 0 & 0 \\ d & 0 & 0 & 0 & 1 \end{array}$$

How to compute the rank distance of two genomes?

M_1:



M_2:

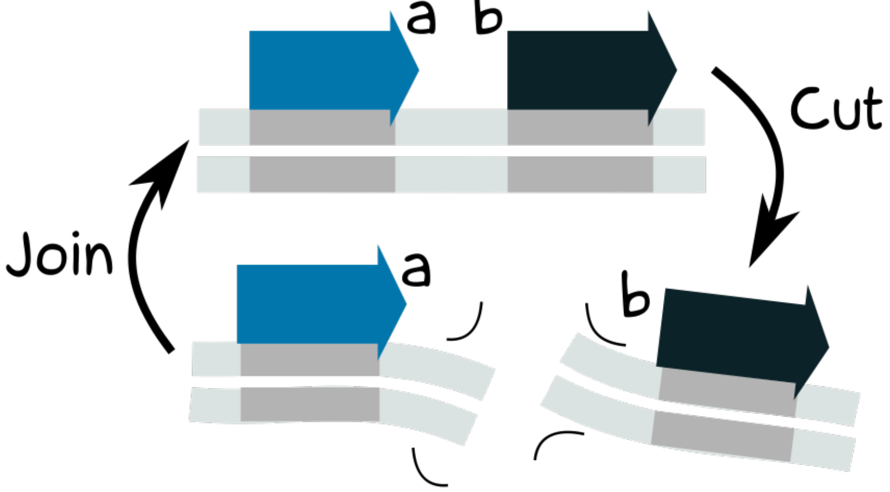


$$d(M_1, M_2) = \text{rank}\left(\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 0 & 1 & 0 \\ b & 0 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & 0 \\ d & 0 & 1 & 0 & 0 \end{array} - \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 0 & 1 & 0 \\ b & 0 & 1 & 0 & 0 \\ c & 1 & 0 & 0 & 0 \\ d & 0 & 0 & 0 & 1 \end{array}\right)$$

$$= \text{rank}\left(\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 0 & 0 & 0 \\ b & 0 & -1 & 0 & 1 \\ c & 0 & 0 & 0 & 0 \\ d & 0 & 1 & 0 & -1 \end{array}\right) = 1$$

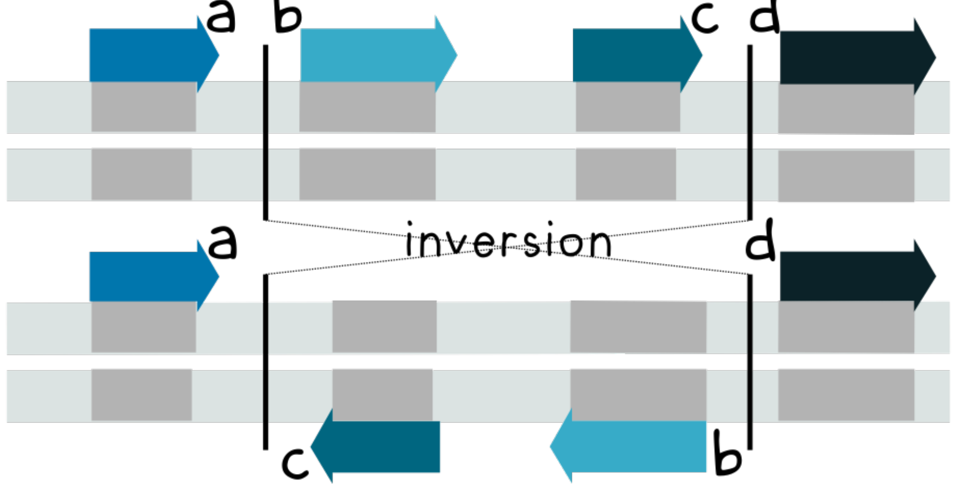
Applications of the rank distance to genome evolution

Rank distance can also be defined as the minimum number of cuts, joins, and double swaps, with weights 1, 1, and 2



(a, b) ⟷ (a) and (b)

Double swap

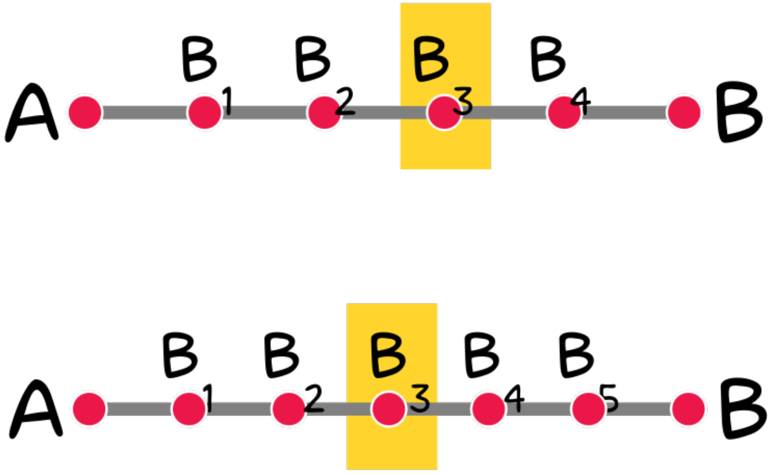


(a, b) and (c, d) ↗ (a, c) and (b, d)

↘ (a, d) and (c, b)

Center conjecture:

Is there always a genome exactly in the middle?

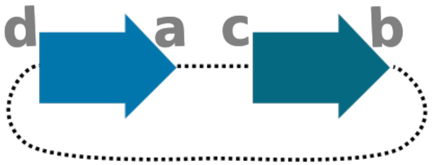


$$\max(d(A,C), d(B,C)) = \left\lceil \frac{d(A, B)}{2} \right\rceil$$

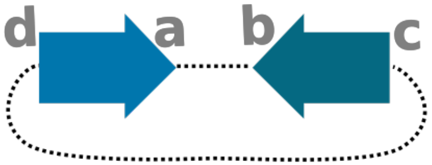
Center conjecture (counterexample):

Is there always a genome exactly in the middle?

A:



B:

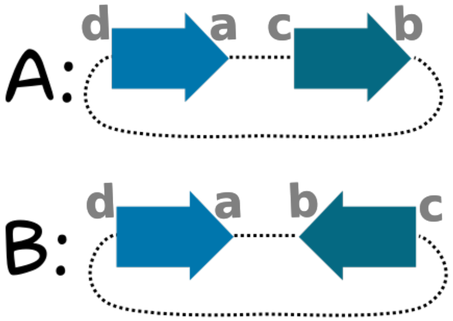


$$\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 0 & 1 & 0 \\ b & 0 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & 0 \\ d & 0 & 1 & 0 & 0 \end{array} \qquad \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 1 & 0 & 0 \\ b & 1 & 0 & 0 & 0 \\ c & 0 & 0 & 0 & 1 \\ d & 0 & 0 & 1 & 0 \end{array}$$

$$d(A, B) = \operatorname{rank}\left(\begin{array}{ccccc} & a & b & c & d \\ a & 0 & -1 & 1 & 0 \\ b & -1 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & -1 \\ d & 0 & 1 & -1 & 0 \end{array}\right) = 2$$

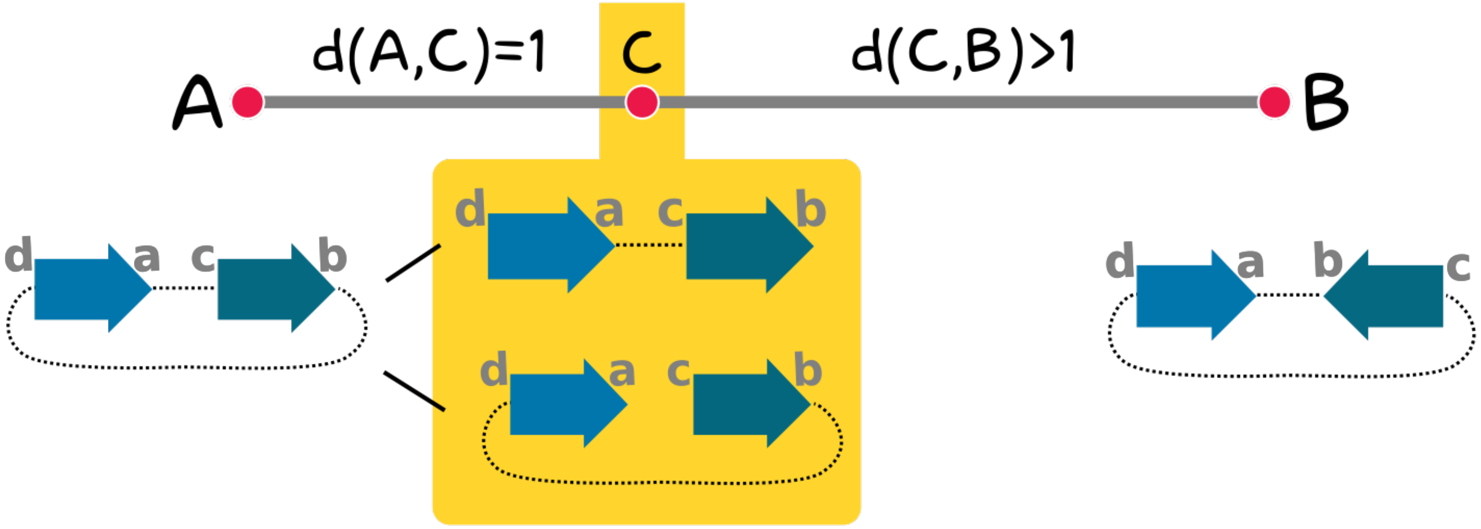
Center conjecture (counterexample):

Is there always a genome exactly in the middle?



d(A, B) = rank(A-B) = 2

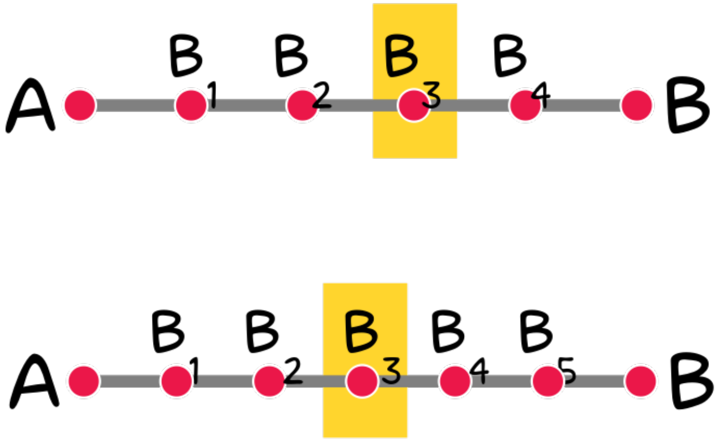
There is no genome with distance 1 from both A and B:



In this example: $\xcancel{\max(d(A,C), d(B,C)) = \left\lceil \frac{d(A, B)}{2} \right\rceil}$

Our goal:

Determine in which cases the center genome is exactly in the middle and, when not in the middle, how far it will be from it.



$$\max(d(A,C), d(B,C)) = \left\lceil \frac{d(A, B)}{2} \right\rceil$$

Our goal: Where is the center genome?

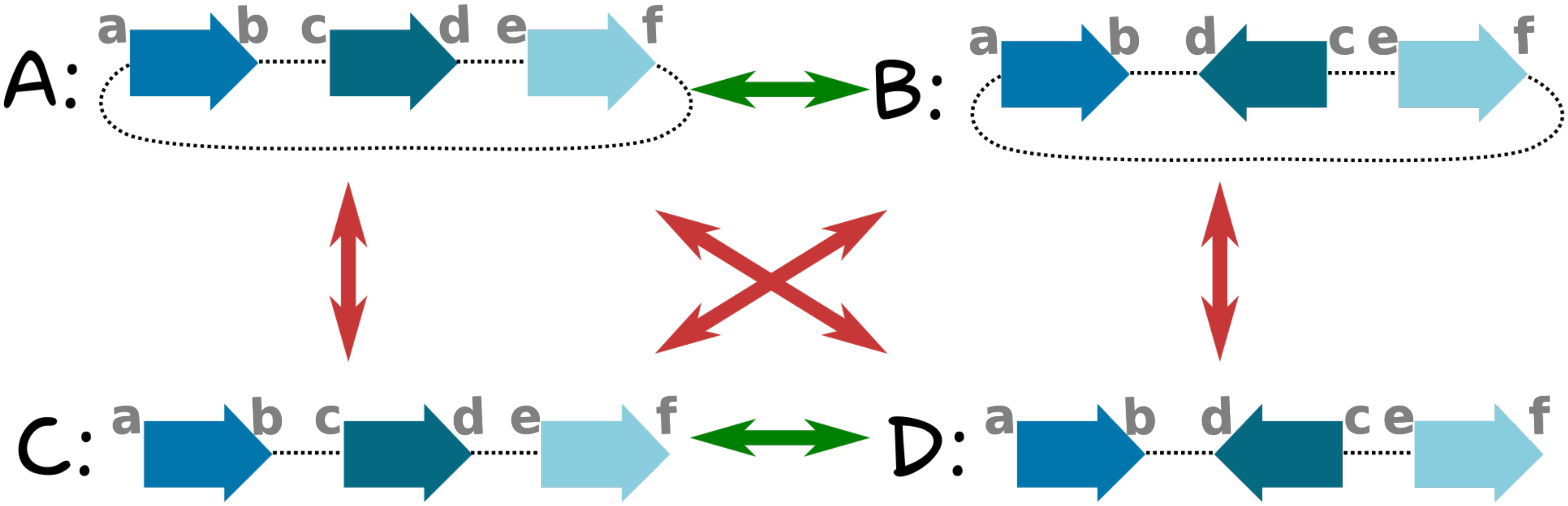
Given a pair of genomes, there are two cases to consider:

Co-tailed genomes

exactly the same telomeres

Not co-tailed genomes

different telomeres



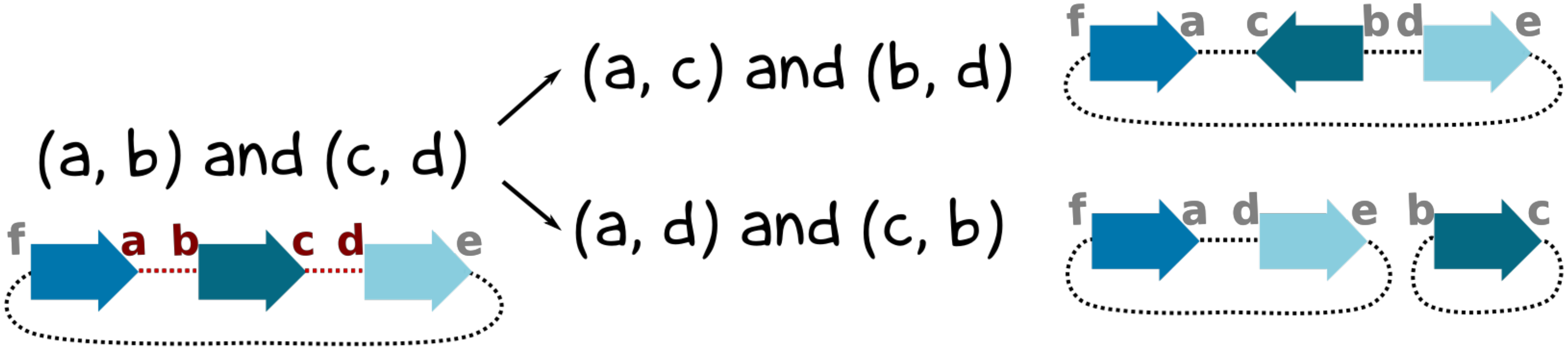
Our goal: Where is the center genome?

Leonid Chindelevitch et al. (2018)

Co-tailed case (same telomeres)

Only double swaps occur in optimal sorting scenarios of co-tailed genomes!

Double swap (weight = 2)



(a, b) and (c, d)

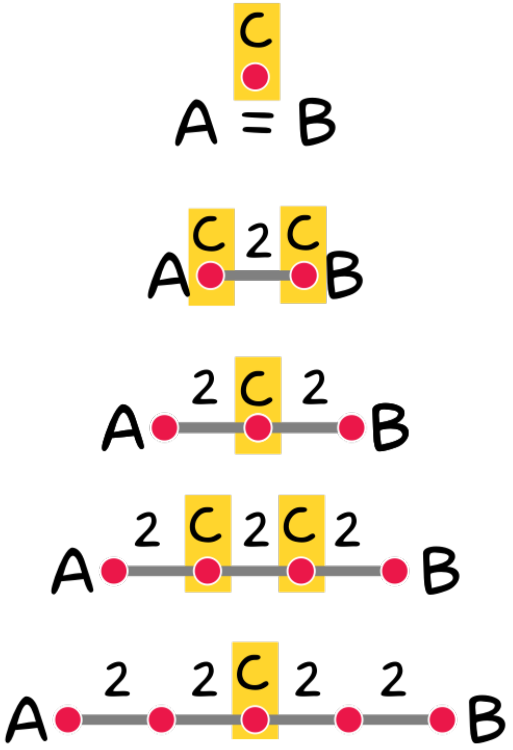
(a, c) and (b, d)

(a, d) and (c, b)

Our goal: Where is the center genome?

Co-tailed case (same telomeres)

As a double swap has weight 2, the distance is even!



d(A,B)	Center genome (C)
0	LB
2	LB + 1
4	LB
6	LB + 1
8	LB

Lower bound (LB):

$$\max(d(A,C), d(B,C)) = \left\lceil \frac{d(A, B)}{2} \right\rceil$$

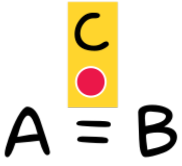
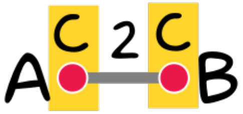
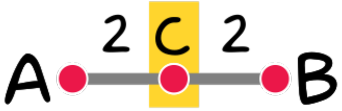
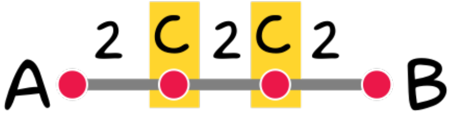
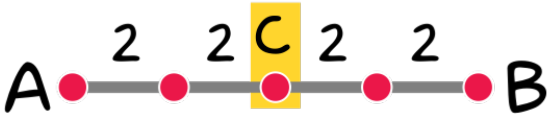
Our goal: Where is the center genome?

Co-tailed case (same telomeres)

d(A,B) is a multiple of 4 : center genome reaches LB!

Otherwise: LB + 1

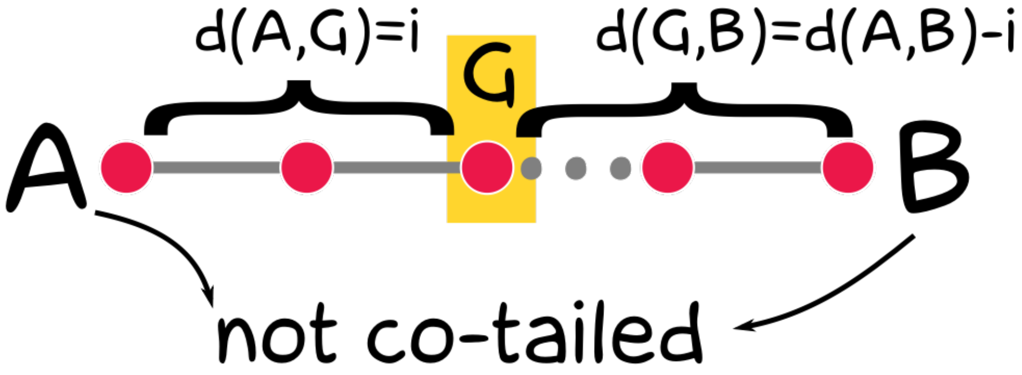
✓

d(A,B)		Center genome (C)
0		LB
2		LB + 1
4		LB
6		LB + 1
8		LB

Our goal: Where is the center genome?

Not co-tailed case (different telomeres)

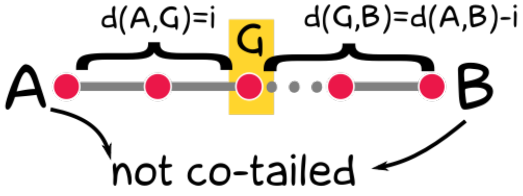
At every step $0 \leqslant i \leqslant d(A,B)$ there is a genome G such that $d(A,G)=i$ and $d(G,B)=d(A,B)-i$...



...so there is a center genome that reaches the LB!

Our goal: Where is the center genome?

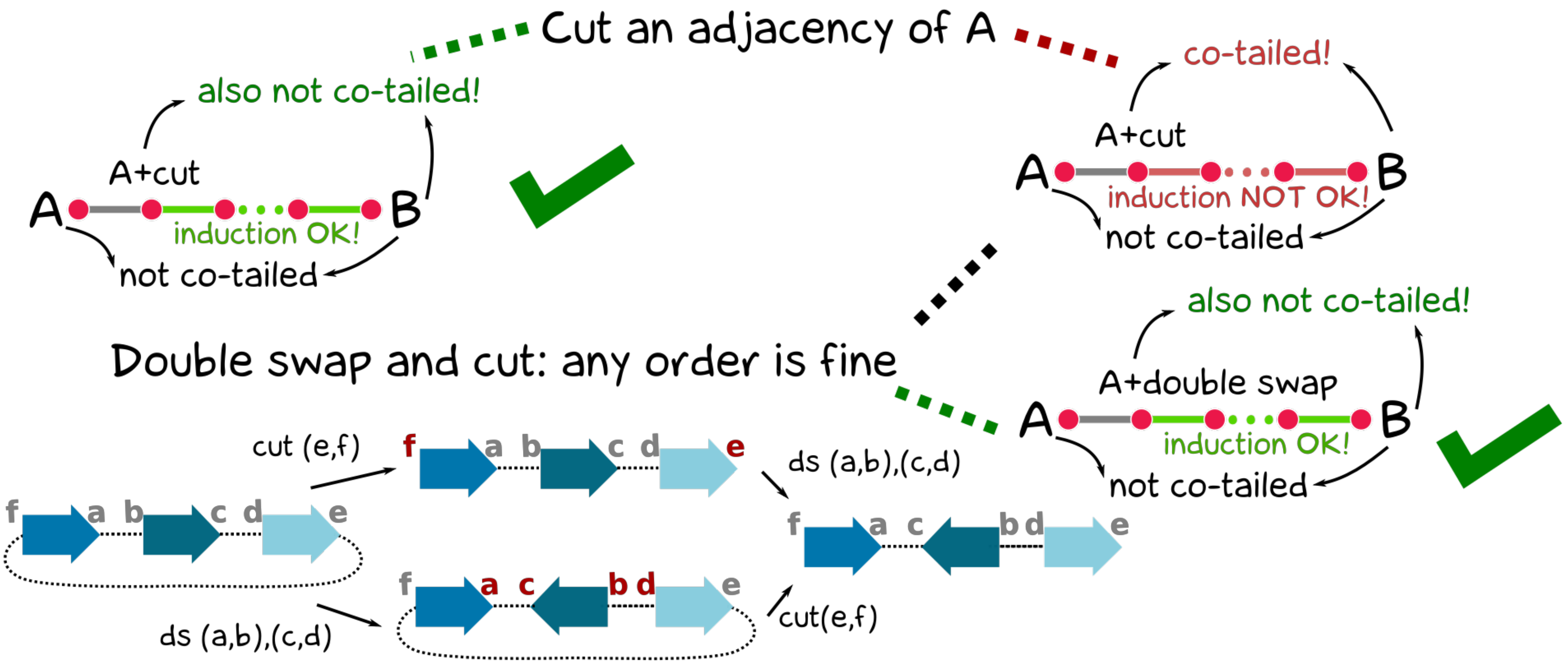
Not co-tailed case (different telomeres)



If A and B are not co-tailed, then there is a genome at every step between 0 and d(A,B)

Proof by induction (idea):

Not co-tailed > It needs a cut!

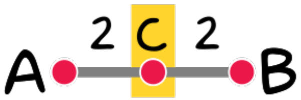

Cut an adjacency of A

Double swap and cut: any order is fine

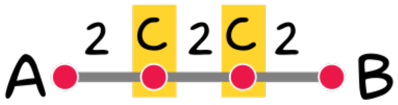
Summary: Where is the center genome?

Center genome (C)

(1) Co-tailed case (same telomeres)



d(A,B) is a multiple of 4: **LB**



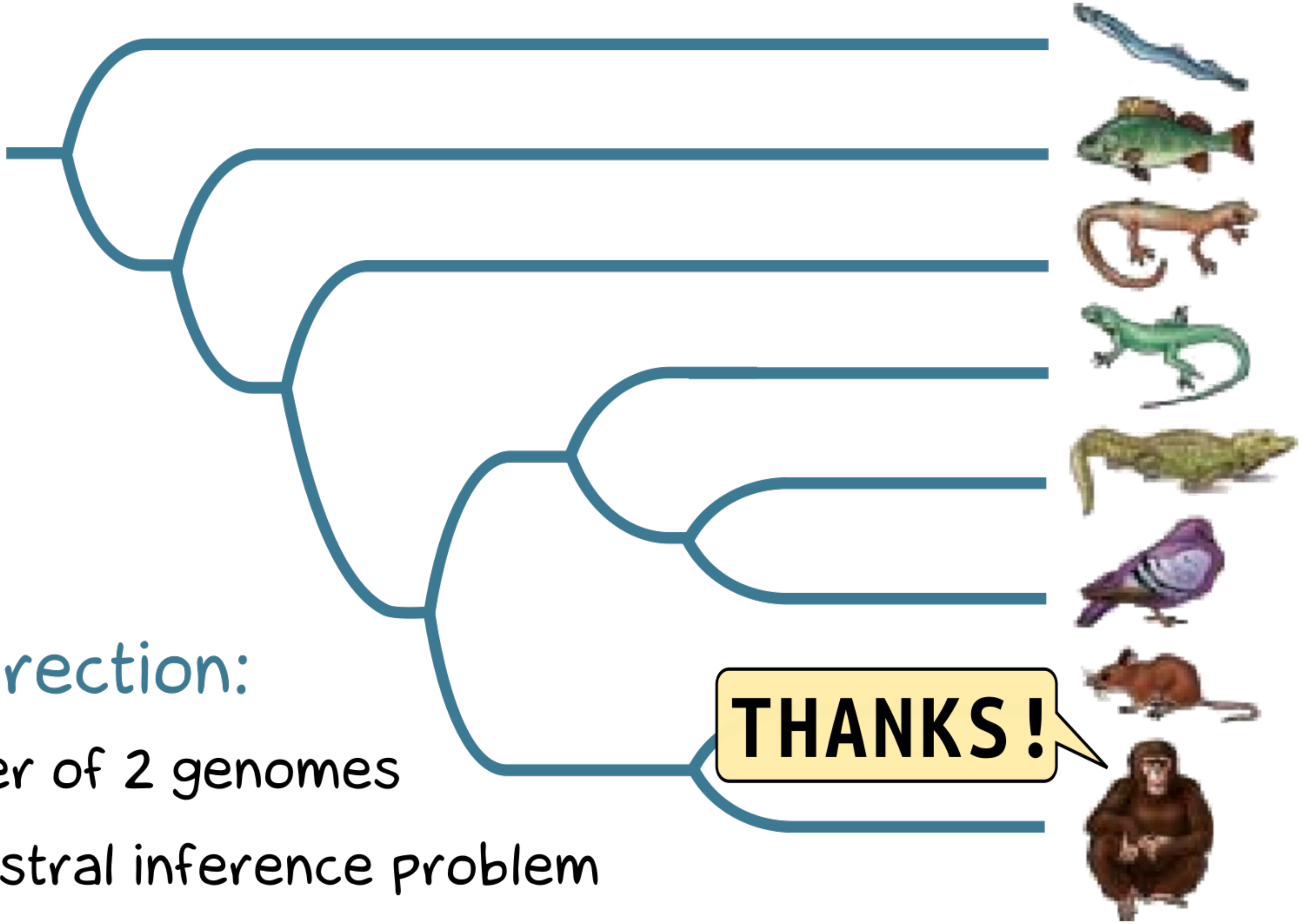
d(A,B) is not a multiple of 4: **LB+1**

(2) Not co-tailed case (different telomeres) **LB**

Lower bound (LB):

$$\max(d(A,C), d(B,C)) = \left\lceil \frac{d(A, B)}{2} \right\rceil$$

Center genome with respect to the rank distance



Future direction:

Apply center of 2 genomes

to the ancestral inference problem