

Algorithms in Comparative Genomics

Universität Bielefeld, WS 2020/2021

Dr. Marília D. V. Braga · Leonard Bohnenkämper

<https://gi.cebitec.uni-bielefeld.de/teaching/2020winter/cg>

Exercise sheet 3, 12.11.2020

Exercise 1 (Breakpoint double distance)

(6 pts)

Consider the following pairs of sing-dup-canonical genomes:

1. $\mathbb{S} = (3\ 1\bar{2})$ and $\mathbb{D} = [1\bar{2}\ 3\ 3\ 1\ 2]$
2. $\mathbb{S} = (3\ 1\bar{2})$ and $\mathbb{D} = [3\ 1\bar{2}\ 3\bar{2}\ 1]$

For each of the given pairs:

- Compute the breakpoint double distance $d_{\text{BP}}^2(\mathbb{S}, \mathbb{D})$;
- Describe the doubled genome $2 \cdot \mathbb{S}$ and give a matching of the genes of $2 \cdot \mathbb{S}$ and \mathbb{D} .

Exercise 2 (SCJ/breakpoint median)

(5 pts)

1. Given three canonical genomes $\mathbb{C}_1 = [1\ 2\ 3\ 4\ 5]$, $\mathbb{C}_2 = [1\ 2\bar{3}\ 5\ 4]$ and $\mathbb{C}_3 = [2\bar{3}\ 1\ 4\ 5]$:
 - (a) Compute a general SCJ median \mathbb{M}_{SCJ} of \mathbb{C}_1 , \mathbb{C}_2 and \mathbb{C}_3 .
 - (b) Is there another SCJ median of \mathbb{C}_1 , \mathbb{C}_2 and \mathbb{C}_3 that is distinct from \mathbb{M}_{SCJ} ?
(Justify your answer by giving a distinct median or explaining why it does not exist.)
 - (c) Is \mathbb{M}_{SCJ} also a breakpoint median of \mathbb{C}_1 , \mathbb{C}_2 and \mathbb{C}_3 ?
If *no*: Compute a breakpoint median of \mathbb{C}_1 , \mathbb{C}_2 and \mathbb{C}_3 .
If *yes*: Is there another breakpoint median of \mathbb{C}_1 , \mathbb{C}_2 and \mathbb{C}_3 that is distinct from \mathbb{M}_{SCJ} ?
(Justify your answer by giving a distinct median or explaining why it does not exist.)
2. Give an example showing that the general algorithm for computing the SCJ median of three canonical genomes \mathbb{C}_1 , \mathbb{C}_2 and \mathbb{C}_3 does not necessarily compute a breakpoint median of \mathbb{C}_1 , \mathbb{C}_2 and \mathbb{C}_3 .
(Hint: it is enough to consider canonical genomes with three genes.)

Exercise 3 (Breakpoint median)

(6 pts)

1. Given three canonical linear genomes \mathbb{L}_1 , \mathbb{L}_2 and \mathbb{L}_3 , let \mathbb{M} be the general breakpoint median of \mathbb{L}_1 , \mathbb{L}_2 and \mathbb{L}_3 . Let $\mathbb{M}_{\mathbb{L}}$ be a genome obtained by breaking, for each circular chromosome of \mathbb{M} , one adjacency with smallest weight into two telomeres. Give an example showing that the obtained $\mathbb{M}_{\mathbb{L}}$ is not necessarily a linear breakpoint median of \mathbb{L}_1 , \mathbb{L}_2 and \mathbb{L}_3 . (Hint: it is enough to consider canonical genomes with four genes.)
2. Can you design an exact polynomial time algorithm to compute the linear breakpoint median of three canonical linear genomes? Justify your answer by describing the algorithm or explaining why it cannot be designed.
3. Can the algorithm for computing the general breakpoint median of $k = 3$ genomes be generalized for computing the general breakpoint median of $k \geq 3$ genomes? Justify your answer by describing the generalized algorithm or explaining why it cannot be generalized.

Exercise 4 (SCJ guided halving)

(3 pts)

The SCJ Double-Guided Halving is the following problem:

Given a duplicated genome \mathbb{D} and two canonical genomes \mathbb{C}_1 and \mathbb{C}_2 , find a canonical genome \mathbb{H} which minimizes the sum $d_{\text{SCJ}}^2(\mathbb{H}, \mathbb{D}) + d_{\text{SCJ}}(\mathbb{H}, \mathbb{C}_1) + d_{\text{SCJ}}(\mathbb{H}, \mathbb{C}_2)$.

Describe an algorithm to solve this problem.