Algorithms in Comparative Genomics

Universität Bielefeld, WS 2021/2022
Dr. Marília D. V. Braga · Leonard Bohnenkämper

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

Exercise sheet 4, 11.11.2021

Exercise 1 (DCJ model) (6 pts)

Given the following pair of canonical genomes:

$$\mathbb{A}^f = [2\ \bar{1}\ \bar{4}\ \bar{3}]\ \ [5\ 7\ 6\ 9\ 8\ 10] \quad \text{and} \quad \mathbb{B}^f = [1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10]$$

1. Draw the relational graph $RG(\mathbb{A}^f, \mathbb{B}^f)$ (or the adjacency graph $AG(\mathbb{A}^f, \mathbb{B}^f)$) and compute the DCJ distance between $\mathbb{A}^f$ and $\mathbb{B}^f$.
2. Give an optimal DCJ sorting scenario from $\mathbb{A}^f$ to $\mathbb{B}^f$. Name the operations in your sorting scenario.
3. If your optimal sorting scenario contains circular chromosomes in two consecutive intermediate genomes, find an alternative optimal scenario in which each circular intermediate chromosome is immediately reintegrated (no need to worry about running time of your procedure).
4. Can you find yet another alternative optimal scenario without circular intermediates?

Exercise 2 (Solution space of DCJ sorting) (6 pts)

1. Given genomes $\mathbb{A}^f = [1]\ \ [4\ 3\ 2\ 5]$ and $\mathbb{B}^f = [1\ 2\ 3\ 4\ 5]$, how many different optimal DCJ scenarios sorting $\mathbb{A}^f$ into $\mathbb{B}^f$ can you find?
2. Given two canonical genomes $\mathbb{G}_1^f$ and $\mathbb{G}_2^f$, let c_k be a k-cycle in $RG(\mathbb{G}_1^f, \mathbb{G}_2^f)$ (with $k \geq 4$), and let $E_\Gamma(\mathbb{G}_1^f, c_k)$ be the set of edges corresponding to adjacencies of genome $\mathbb{G}_1^f$ that are in c_k.

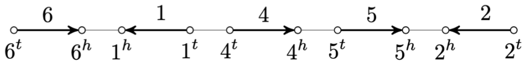
   What is the number of distinct DCJ operations that modify genome $\mathbb{G}_1^f$ and split c_k into two cycles?

   (*Tip: For each pair of edges in $E_\Gamma(\mathbb{G}_1^f, c_k)$, there is exactly one DCJ splitting c_k into two cycles.*)

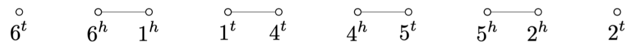
Exercise 3 (Combinatorial problem) (6 pts)

A genome can be seen as a matching on the set of all $2n$ extremities, where adjacencies correspond to two matched extremities, and telomeres are extremities that are not matched to any other.

For instance, the genome



corresponds to the matching



1. What is the total number of different circular genomes with n genes? Try to find a recursive equation to express the number C_n of circular genomes of n genes.

   (*Tip: a circular genome does not have telomeres, only adjacencies, therefore it is the same as counting all perfect matchings, that is, matchings where all vertices are matched.*)
2. The total number of genomes with n genes can be obtained with the following approach: consider all possibilities of dividing the $2n$ vertices into two bags:
   - Telomeres, with $2k$ vertices for $(k = 0, 1, 2, \ldots, n)$; and
   - Adjacencies, with $2(n-k)$ vertices, for which the same formula from item [1.] applies.

   Let $G_{n,k}$ denote the number of genomes with n genes and $2k$ telomeres. Note that $G_{n,0} = C_n$. The total number of genomes with n genes can then be expressed as

   $$G_n = \sum_{k=0,1,2,\ldots n} G_{n,k}\ .$$

   Try to find a formula for computing $G_{n,k}$.