Algorithms in Comparative Genomics

Universität Bielefeld, SS 2024
Dr. Marília D. V. Braga · Dr. Roland Wittler · M.Sc. Leonard Bohnenkämper

https://gi.cebitec.uni-bielefeld.de/teaching/2024summer/cg

Exercise sheet 10, 14.06.2024

Exercise 1 (Common intervals on permutations) (5 pts)
Recall that a genome g contains a common interval c if all genes in c appear consecutively in g, and that $GS(C)$ defines the set of all genomes that contain all gene clusters (in this case common intervals) $c \in C$. Recall further the relation to the *Consecutive Ones Problem*.

Consider permutations on the set of genes $\{0, \dots, 14\}$ and the following set C of common intervals:

$$C := \{\{0,1,2,3,4,5\}, \{3,4\}, \{4,5,6,7,8\}, \{9,10,11,12,13\}, \{10,11,12\}, \{11,12,13\}\}$$

1. Inform yourself about PQ-trees, and use the applet provided on the lecture website to construct the PQ-tree corresponding to C.
2. How many (linear) genomes are in $GS(C)$? Can you derive a general formula?
3. Which other further common intervals that are not in C are implied by C?
4. Which framed common intervals and nested common intervals can you find?

Exercise 2 (Framed common intervals on permutations) (2 pts)
Decompose the framed common interval $[2\,\{1,3,4\}\,-5]$ over the set of genes $\{1, \dots, N\}$ into an equivalent set of common intervals over the set $\{1^h, 1^t, \dots, N^h, N^t\}$.

Exercise 3 (Unsigned adjacencies on sequences) (3 pts)
Consider the multiset of genes $\{1,2,3,3,4,5,6,6,7\}$ and sketch the Euler graph for the following set of unsigned adjacencies including auxiliary node $\circledast$:

$$\big\{\{1,2\}, \{1,3\}, \{2,3\}, \{4,6\}, \{4,7\}, \{5,6\}, \{6,7\}\big\}$$

Verify consistency of the set of adjacencies.

Exercise 4 (Minimal conflicting subsets) (3 pts)
Devise an algorithm in pseudo code to determine all minimal conflicting subsets of a given set of gene clusters. You can make use of the function $GS(C)$ to get the set of all genomes that contain all gene clusters $c \in C$.

Hint: Have a look at the pages 63–76 of the slides of the lecture.

Exercise 5 (Reconstruction of ancestral gene clusters) (3 pts)
In the lecture, we discussed the objective of finding, under all consistent labelings, a labeling of minimal parsimony weight. Recall the visualization of the search space, where the y-axis shows the parsimony weight, and moving along points on the plane corresponds to removing gene clusters from a labeling. (See pages 40 or 58 of the slides of the lecture. Note that this diagram does not include an x-axis.) Recall further that after the removal of a cluster from a node, a re-optimization is performed, i.e., the same cluster is removed from further nodes to minimize the weight. An arrow in the visualization includes this re-optimization.
Do horizontal arrows make sense, i.e., can there be scenarios of presence/absence of a gene cluster for which the removal of the cluster from a node followed by re-optimization does not lead to an increase of the weight?