Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye

Algorithms in Comparative Genomics
Summer 2018

Exercises

Number 7, return 2018 June 29

1. Consider the following computational problem: Given three genomes A, B, C and a distance metric d (defining pairwise distances $d_{A,B}$, $d_{A,C}$ and $d_{B,C}$), find a median genome M, i.e. a genome
$$M := \operatorname*{argmin}_{G \in \{\text{all genomes}\}} \{d_{A,G} + d_{B,G} + d_{C,G}\}.$$
   (a) Show that choosing the minimum of $d_{A,B} + d_{A,C}$, $d_{A,B} + d_{B,C}$, and $d_{A,C} + d_{B,C}$ (i.e. the minimum spanning tree) gives a $\frac{4}{3}$-approximation to the optimal median distance.
   (b) Show that $\frac{1}{2}(d_{A,B} + d_{B,C} + d_{A,C})$ is a lower bound for the median-of-three problem.
   (c) Show that if the largest of the three pairwise distances is equal to the sum of the other two, the solution to (a) gives an optimum.

2. List all DCJ median genomes of $A = [1, -2, -3]$ and $B = [1, 2, 3]$.

3. (a) Download the rococo software from `https://bibiserv.cebitec.uni-bielefeld.de/rococo` and the two data files **phylogenetic tree** and **gene order data** from the class web page `http://gi.cebitec.uni-bielefeld.de/teaching/2018summer/cg`.
   Run rococo (unsigned or signed adjacencies) on this data set. View the results with the Rococo-Compare Tool `comp`, also downloadable from the rococo web site.
   Look for the cluster that contains the four genes **nrdl**, **nrdH**, **rpmJ** and **nadE**. What happened with this cluster in *C. urealyticum*? What can you say about the function of the genes in this cluster, in particular the gene with the number 1306?
   (b) The gene cluster studied in part (a) has a particular structure: A set of genes with a particular "extra gene" in a small subset of the genomes. Can you think of other gene cluster patterns that might be evolutionarily interesting?