Exercises – Phylogenetics

Universität Bielefeld, SS 2019
Dr. Roland Wittler, M. Sc. Tizian Schulz
https://gi.cebitec.uni-bielefeld.de/Teaching/2019summer/Phylogenetik

Exercise Sheet 5 — 09.05.2019

Due: 16.05.2019

Task 1 DNA Grid Graph. **(2 points)**
What is the number of edges $|E|$ for a DNA grid graph $G = (V, E)$ for sequence length m? Derive a formula for $|E|$ that only depends on m and explain it.

Task 2 Spanning Tree Heuristic. **(4 points)**
Search for a *most parsimonious tree* of the taxa A to E with regard to the following sequences.

A: T C G T T
B: A A T T T
C: T C T G T
D: T C T T G
E: A A T A T

Let G be a *DNA grid graph* that contains all sequences of length 5 and therefore particularly the nodes that correspond to the taxa A to E. Use the spanning tree heuristic to approximate a *Steiner tree* for the nodes. Proceed like this:

Step 1: Shortest paths. Calculate all pairwise Hamming distances for the given taxa A to E and create the edge-weighted graph G'.

Step 2: Spanning tree. Create a minimum spanning tree T' in G' using Prim's algorithm starting at the node representing sequence C. Write down the order of edges you have chosen. To avoid different results in this step, add the lexicographically smallest edge to T' each time you have a choice. (Hint: We do consider undirected edges here which means, e.g., $(A, B) < (A, C) = (C, A) < (B, C)$)

Step 3: Map back to G. Draw the part of the grid graph G that contains T'. Add all sequences as nodes into the tree such that the Hamming distance between all nodes is exactly 1. Try to add as few nodes as possible and reuse some nodes for different edges.

Task 3 Spanning Tree Heuristic: 2-Approximation. **(3 points)**

We have shown that the approximation factor of the spanning tree heuristic is at most 2. Use the given graph (with terminal nodes $v_1, ..., v_n$) to show that this value is *tight*, i.e., the approximation factor is not smaller than 2.