Bielefeld University
Faculty of Technology

AG Genominformatik
Dr. Marília D. V. Braga
Dr. Daniel Doerr

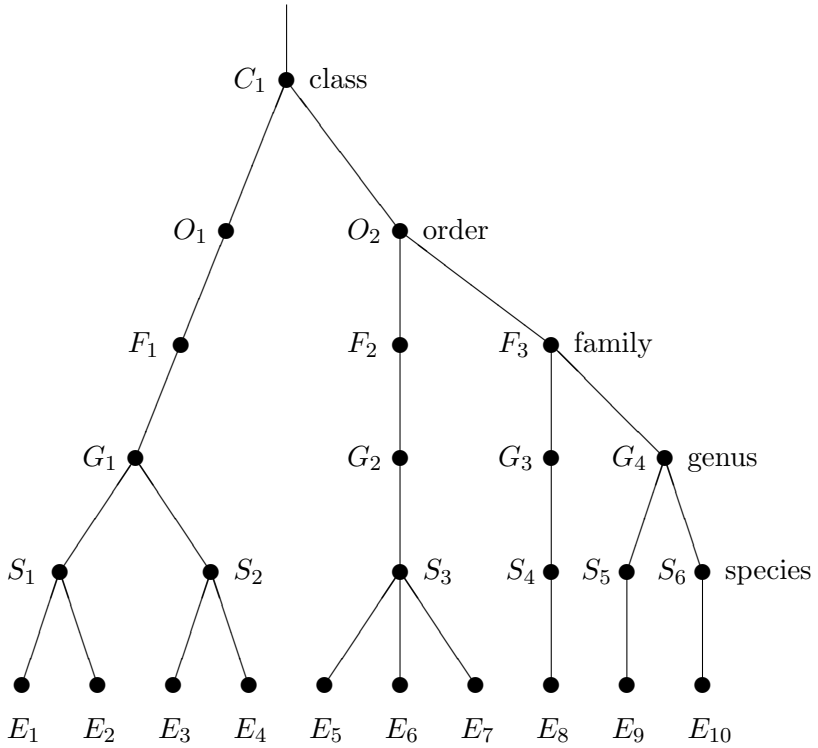
Algorithms in Genome Research

Winter 2019/2020

Exercises

Number 7, Discussion: 2020 January 10

Assume we have given 5 metagenomic reads $R_1, \ldots, R_6$ and a database containing 10 entries $E_1, \ldots, E_{10}$ from 6 different species $S_1, \ldots, S_6$. The following taxonomy is given for the genomes and the contained sequences:



Let the following be the BLAST output as pairs (entry, bit-score), when the reads are BLASTed against the database:

$R_1 : (E_8, 50)(E_{10}, 47)(E_6, 43)(E_9, 42)(E_3, 42)(E_5, 34)(E_1, 33)(E_2, 27)(E_7, 24)(E_4, 23)$
$R_2 : (E_7, 74)(E_5, 69)(E_9, 64)(E_6, 63)(E_{10}, 61)(E_4, 61)(E_1, 59)(E_2, 54)(E_8, 34)(E_3, 33)$
$R_3 : (E_3, 60)(E_4, 60)(E_1, 58)(E_2, 55)(E_7, 53)(E_8, 51)(E_6, 50)(E_9, 47)(E_5, 45)(E_{10}, 38)$
$R_4 : (E_5, 37)(E_4, 35)(E_6, 34)(E_7, 34)(E_{10}, 33)(E_1, 28)(E_8, 27)(E_2, 25)(E_9, 23)(E_3, 18)$
$R_5 : (E_2, 87)(E_4, 41)(E_1, 40)(E_3, 39)(E_6, 39)(E_8, 37)(E_{10}, 33)(E_9, 32)(E_7, 30)(E_5, 29)$

1. What would be the predictions by MG-RAST?
2. What would be the predictions by MEGAN with threshold 10%?

(please turn over)

3. Assume another metagenomic read R_6 for which entry E_{10} yields the best BLAST result and the BLAST hits of entries $E_1, E_5, E_6, E_7, E_8, E_9$ (and E_{10}) have a score above 90% of that best score. Both SOrt-ITEMS and CARMA3 now BLAST E_{10} against the database $\{R_6, E_1, E_5, E_6, E_7, E_8, E_9, E_{10}\}$. Assume the following output:

   $E_9 : (E_{10}, 140), (E_8, 130), (E_7, 120), (E_9, 100), (E_6, 90), (E_5, 80), (E_1, 50)$ and (R_6, s)

   The taxonomic assignment of R_6 depends on score s. Separate all possible values of s into appropriate ranges and specify the corresponding assignment by (a) SOrt-ITEMS and (b) CARMA3. (For SOrt-ITEMS, assume that all entries have been assigned to rank *genus*. Feel free to ignore borderline cases such as $s = 50, 80, \ldots$)