**Lab Project GP AGO 2020/2021 v1.00**

Create and test a phylogenetic pipeline to infer a species/genome tree from a set of genomes by clustering, inferring gene families and their trees. Alternative approaches are welcomed but require approval.

**Part I.**

Choose a set of species of a reasonable size. Examples:

- minimum one hundred similar viruses (e.g., coronaviruses),
- minimum 30 bacterial or archaeal species having up to 5000 genes,
- minimum 10 more complex species; here, I suggest limiting the number of genes,
- a mixture of species from different kingdoms, please contact us to approve the choice.

Download from a public database (e.g., NCBI) proteomes of the corresponding species.

**Part II. Sequence clustering.**

Cluster the sequences using, e.g., BLAST (pairwise comparison) + MCL or alternative methods. Faster tools (see MMSeqs2) are recommended in the case of large datasets.

**Part III. Gene families inference (one-to-one case)**

Create gene families from each cluster. Replace gene identifiers with the name/id of the corresponding species.
Depending on the further choices of additional cluster filtering may be needed, e.g.,

- removing clusters (e.g., small or having sequences from one species only)
- removing sequences from clusters to obtain one-to-one correspondence between genes and species.

**Part IV. Multi-alignments**

Compute alignments for every gene family.

**Part V. Gene family trees**

Infer phylogenetic trees from each multialignment using NJ, ML or MP methods.
Using the trees from the alignment tool is not recommended (-13 points).

**Part VI. Species tree inference**

Compute the final tree species using consensus and supertree methods.

**Part VII. Analysis of the results**
Compare your final tree with the known hypotheses on the evolution of your set of species. Also, check the NCBI taxonomy and the tree from timetree.org (if applicable). Write a report with the species' description, a summary of used methods with runtime (including bottlenecks), and biological conclusions.

Additional steps with more scoring options:

**Part V.a. Gene trees filtering** (+4 points)

For every gene family, by using bootstrapping, eliminate weekly supported trees. Check if this approach gives a better species tree than the tree without filtering.

**Part V.b. Paralogous gene family trees** (+4 points)

Use clusters without the removal of sequences, i.e., allow paralogs. Then, compute the supertree. Compare the results with the tree obtained from "one-to-one" clusters.

**Additional information**

Send using moodle:

1. Scripts (bash, python, etc.) - the pipeline should be automated and parallel (if possible).
2. The report with results (odt, doc, pdf).
3. Gene trees (one file) + species tree (one file) in a newick format.
4. In case of more variants covered, e.g., V.a, attach more newick files.
5. README - how to use scripts + info on the attached newick files.

Please **do not submit** your sequence data files.

Scoring rules:

1. Max is 40 points = 27 base + 2x4 bonus points + 5 project presentation on the 27th of January, 2021 (a strict deadline!) during the last lecture.
2. The project presentation during the last lecture is approx. 10-15 min; the results should be submitted one day before the last lecture.
3. In case of submissions without the presentation on 27/01/2021, it is still required to present the project with standard I and II session deadlines. In such a case, the presentation is only with the corresponding lab teacher.

8/1/21 Update: project presentation on the 27th of January, 2021 (previously incorrect date).