

Genomika Porównawcza - Laboratorium 7: Consensus trees

Barbara Pawłowska

The objective of this study was to reconstruct and analyze the phylogeny among nine yeast species from the Genolevures project [1]. The dataset comprised 1690 gene families, from which 500 families were randomly selected, each containing nine sequences. For each family, multiple sequence alignment was performed using the MAFFT algorithm, followed by phylogenetic tree reconstruction with IQ-TREE2. As a result, 500 phylogenetic trees were obtained.

To determine the consensus tree, the --bcon option in IQ-TREE2 was employed, based on the previously constructed set of 500 trees. This method clusters identical tree topologies, and the calculated percentage values represent the proportion of trees supporting each specific branch. Since trees generated by IQ-TREE2 are unrooted, the consensus tree was initially preserved in its unrooted form. To facilitate comparison with the rooted reference tree, the consensus tree was subsequently rooted along the branch corresponding to the species YALI.

The reference topology for the species from the Genolevures project [1] is:

(YALI,(DEHA,(((SAKL,KLTH),(KLLA,ERGO)),(ZYRO,(SACE,CAGL)))))

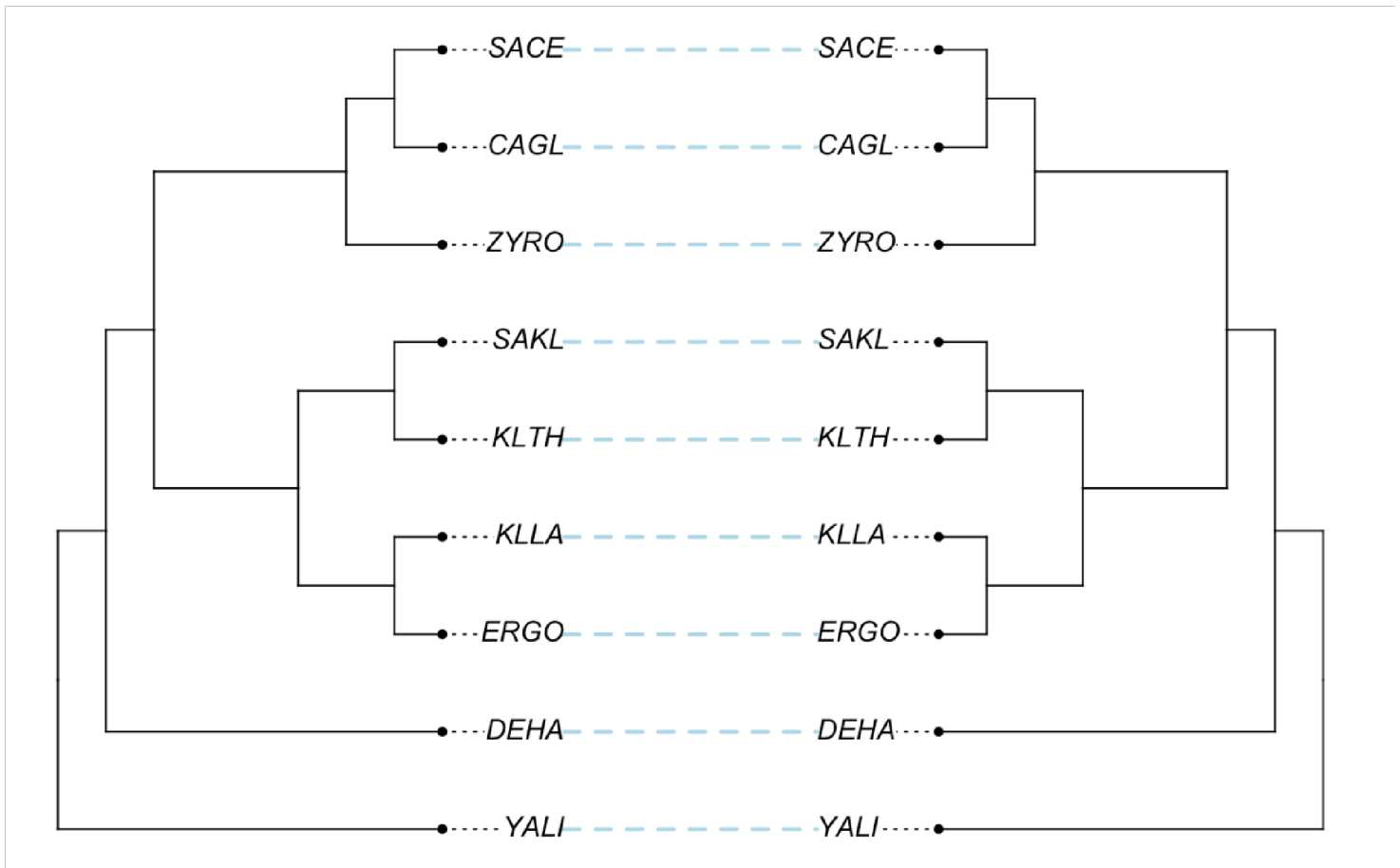
The consensus tree obtained in this study has the following topology:

(YALI,(DEHA,(((ERGO,KLLA)40,(KLTH,SAKL)44)30,(ZYRO,(CAGL,SACE))53)70)91)

Following the rooting of the consensus tree, the Robinson–Foulds (RF) distance was calculated to assess the topological congruence between the trees. The obtained RF value of 0 indicates complete topological agreement between both trees. This demonstrates that, despite differences resulting from the order of branch notation in the Newick format, the branching pattern of the consensus tree accurately reflects the topology of the reference tree.

It is important to emphasize that only 500 out of the 1690 available gene families were utilized in this analysis. As different genes may reflect distinct evolutionary histories, the selection of a specific subset can influence the results, including bootstrap values and the level of support for individual branches. Nevertheless, despite potential variation introduced by the chosen subset of gene families, the consensus tree successfully reconstructed the reference topology, as confirmed by the RF value of 0.

REFERENCE TREE VS CONSENSUS TREE



[1] Sherman, D. J., Martin, T., Nikolski, M., Cayla, C., Souciet, J. L., Durrens, P., & Génolevures Consortium. (2009). Génolevures: Protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Research*, 37(Database issue), D550–D554.