Phylogenetic pipeline project

Kornel Natoński

January 2026

Abstract

In this work, I describe the phylogenetic pipeline I developed to produce species trees for the genus *Bifidobacterium*. My pipeline uses MMseqs2 to cluster sequences into gene families, MAFFT to align sequences, and FastTree to build individual gene trees. To combine them into a final species tree I used IQ-TREE for consensus trees and ASTER package for supertrees. The obtained trees show moderate success in replicating the overall tree structure. The supertrees managed to quite accurately reflect the key groups described in the original paper [1].

1 Introduction

Bifidobacterium is a genus of bacteria present in intestinal tracts of mammals. They are often used in probiotics.

In this work I tried to reproduce the species tree of *Bifidobacterium* obtained in a 2018 article *Tracking the Taxonomy of the Genus Bifidobacterium Based on a Phylogenomic Approach* [1] presented on Figure 1. They produced a tree with clear clustering of species into groups corresponding to their ecological hosts (human-associated *adolescentis*, insect-associated *asteroides* group etc.). I tried to get all the same species and strains that they used for my dataset. I managed to do that except for one. *Bifidobacterium saeculare* is missing because it has been reclassified as a subspecies of *Bifidobacterium gallinarum* and was unlisted from the NCBI database [2]. In the end I used 55 species, for each of them taking the strain described as type strain in the original's supplementary materials, so the same one that they used.

2 Methods

2.1 Data preprocessing

2.2 Clustering

First I gathered all the needed NCBI accessions of the strains of interest in *accessions.csv*. After gathering all of them and verifying they're the strains I need I used MMseqs2 to cluster them into gene families. I used the Entropy

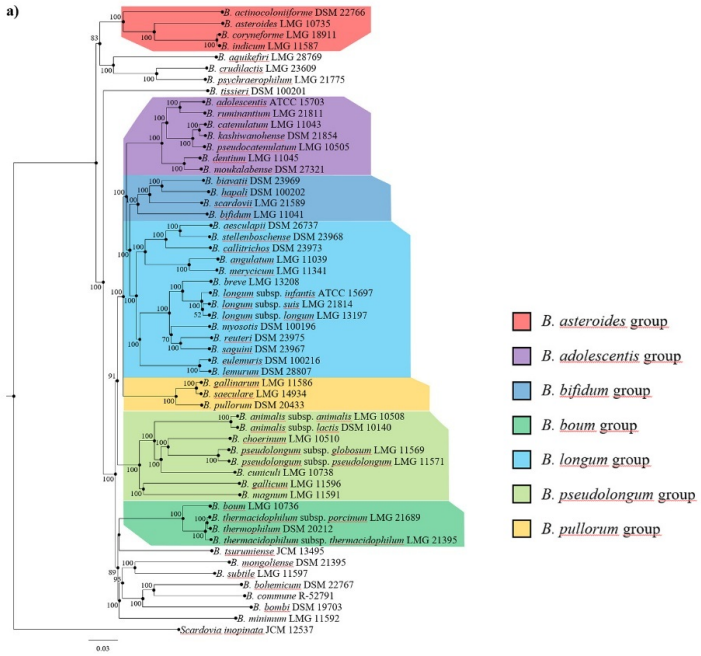

Figure 1: Reproduction of the species tree of *Bifidobacterium* from *Tracking the Taxonomy of the Genus Bifidobacterium Based on a Phylogenomic Approach* [1].

computer cluster to perform my computations. Thanks to that the clustering only took 1min 11s, thanks to utilizing 40 CPU cores. I removed paralogs from the families leaving behind a copy with paralogs for later use.

2.3 Alignment

Then I aligned the sequences using MAFFT. Both for the gene families with paralogs removed (*filtrered_gene_families*) and ones with paralogs present (*raw_gene_families*). I used *run_align_inplace.py* script submitting the job with *align_job.sh*. This process took 27min 16s utilizing 56 CPU cores.

2.4 Gene trees

Computing trees for each gene family with FastTree took 5min 39s on 56 cpu cores. Computations were run togther for *filtered_gene_families* (no paralogs) and *raw_gene_families* (with paralogs), because if there's less then 4 trees in the filtered version of a gene family, we remove the unfiltered version too, since it contains less than 4 unique species. I used *run_trees.py* script submitting the job with *genetrees_job.sh*.

Using *make_map.py* I made a mapping of unique sequence names to the species they belong to. This will be used to make the version of the tree with paralogs.

2.5 Genome Trees

2.5.1 Supertrees

To create my supertrees I used the state-of-the-art ASTER phylogenomic package [4]. First I concatenated my trees into single files (separately for paralogs and no paralogs cases). Then I run the appropriate programs for each case. They're both very fast so it could be performed in an interactive session, just like clustering before. To compute a supertree utilizing the sequences with paralogs removed I used the Astral program from the ASTER package. To compute a supertree utilizing all sequences, without removing paralogs I used the Astral Pro 3 program from the ASTER package.

2.5.2 Consensus trees

For consensus trees I fed the gene trees created by FastTree into methods IQTree package [8]. I can only use the trees containing all 55 species. There turned out to be 24 of them. I produced a majority consensus tree (min. sup. ¿0.5) and a greedy consensus tree implemented as "extended consensus" in IQTree.

3 Results

In the end I obtained 4 genome trees. Their comparison to the literature is shown in 1, where similarity is defined as

$$Similarity = (1 - RF/max\_possible\_RF) * 100\%$$

. Supertrees performed identically to each other, outperforming the consensus trees according to the RF distance metric.

Another means of comparing the trees is visual inspection. Species names were colored the same as in the original paper to highlight the groups it describes. It can be seen that in case of the supertree the highlighted groups were well-preserved with the exception of *Bifidobacterium dentium* splitting away from the *B. adolescentis* group.

Table 1: Comparison of reconstructed trees against the literature.

Tree	**RF distance**	**Similarity (%)**
Greedy consensus	37	64.42
Majority consensus	39	62.50
Supertree (paralogs)	24	76.92
Supertree (no paralogs)	24	76.92

4 Conclusions

The overall performance of my pipeline did not result in any tree very similar to the one found in literature. However the supertree approach managed to produce a tree were the clusters described in the paper are well preserved. This means the tree does reflect the species origin in a biologically significant sense.

There may be several reasons possible improvements to the methodology. Maybe some of the species should be discarded to obtain more gene families that contain all the species. Maybe instead of FastTree the gene trees should also be computed with the more computationally demanding IQtree. Maybe the data selection could be more thorough and despite the lackluster data availability a dataset closer to the one from [1] could be produced.

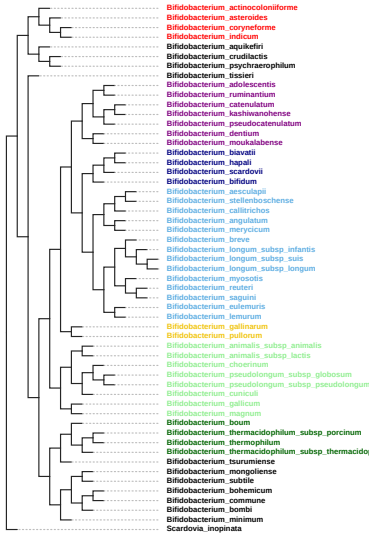

Figure 2: Tree from literature.

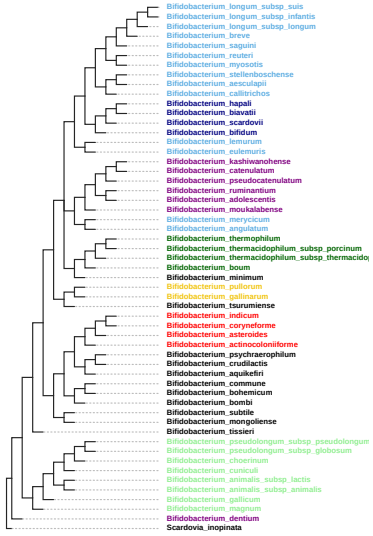

Figure 3: Greedy consensus tree.

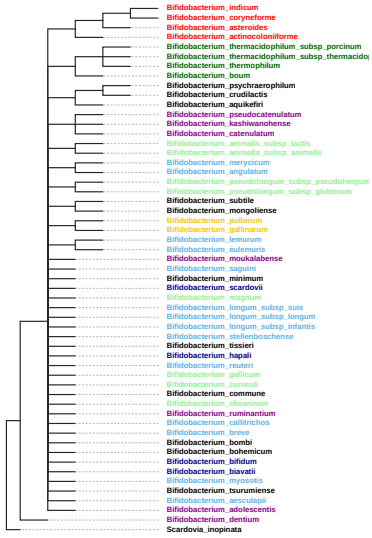

Figure 4: Majority consensus tree.

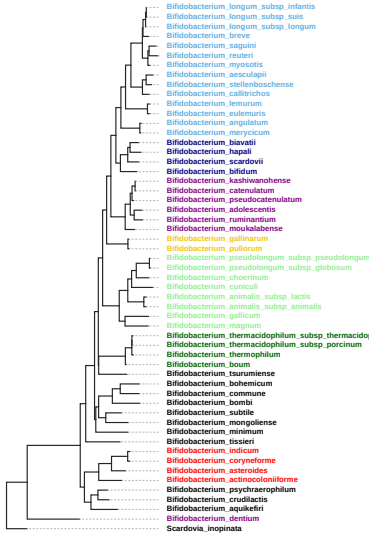

Figure 5: Supertree (no paralogs case identical to paralogs case).

Figure 6: Visual comparison of trees.

References

[1] Lugli, Gabriele Andrea, Christian Milani, Sabrina Duranti, et al. "Tracking the Taxonomy of the Genus Bifidobacterium Based on a Phylogenomic Approach." *Applied and Environmental Microbiology* 84, no. 4 (2018): e02249-17. https://doi.org/10.1128/AEM.02249-17.

[2] Liu, Dan Dan, Hao Wang, and Chun Tao Gu. "Proposal of Bifidobacterium Saeculare Biavati et al. 1992 as a Later Heterotypic Synonym of Bifidobacterium Gallinarum Watabe et al. 1983 and Bifidobacterium Gallinarum Subsp. Saeculare Subsp. Nov." *International Journal of Systematic and Evolutionary Microbiology* 70, no. 11 (2020): 5964–68. https://doi.org/10.1099/ijsem.0.004474.

[3] Steinegger, Martin, and Johannes Söding. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35, no. 11 (2017): 1026–28. https://doi.org/10.1038/nbt.3988.

[4] Zhang, Chao, Rasmus Nielsen, and Siavash Mirarab. "ASTER: A Package for Large-Scale Phylogenomic Reconstructions." *Molecular Biology and Evolution* 42, no. 8 (2025). https://doi.org/10.1093/molbev/msaf172.

[5] Rodriguez-R, Luis M, Santosh Gunturu, William T Harvey, et al. "The Microbial Genomes Atlas (MiGA) Webserver: Taxonomic and Gene Diversity Analysis of Archaea and Bacteria at the Whole Genome Level." *Nucleic Acids Research* 46, no. W1 (2018): W282–88. https://doi.org/10.1093/nar/gky467.

[6] Price, Morgan N., Dehal, Paramvir S., and Arkin, Adam P. "FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix." *Molecular Biology and Evolution* 26, no. 7 (2009): 1641–1650. https://doi.org/10.1093/molbev/msp077.

[7] Katoh, Kazutaka, and Standley, Daron M. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30, no. 4 (2013): 772-80. https://doi.org/10.1093/molbev/mst010.

[8] Nguyen, Lam-Tung, Schmidt, Heiko A., von Haeseler, Arndt, and Minh, Bui Quang. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32, no. 1 (2015): 268-74. https://doi.org/10.1093/molbev/msu300.