Master thesis proposal
Clustering of microbiome data with structural and excess zeroes and small samples

Krzysztof Bartoszek

October 10, 2025

Due to the advancement in current sequencing techniques microflora in individuals is easier to measure and can be used to identify an individual's population. A feature of microbiome data is that most of the bacterial counts are 0. These zeroes can be structural, there are no representatives of such a bacteria. However, other 0s can be due to very low presence or technical errors. Hence, these 0 counts have to be carefully handled when doing clustering based on the microflora, especially when some Gaussian models/approximations are used. Kaul et al. [3, 4] proposed two such methods. Furthermore, if the microbiomes come from different species, then incorporation of the phylogeny is necessary. An approach in this direction has been proposed in [1, 2]. A starting point for this thesis will be the master thesis by Jun Li [5]. The goal of the thesis is to

1. investigate the methods, especially from a phylogenetic context
2. study their behaviour (esp. ability to correctly cluster) for different (esp. small) sample sizes [like 6, have] and signal strength (esp. weak).
3. investigate how inclusion of the phylogeny affects the inference, or what are the consequences of ignoring the phylogeny

For the thesis simulated and from the literature data [6] will be used. Furthermore, Dr. hab. Mioduchowska (Department of Genetics and Biosystematics, Faculty of Biology, University of Gdańsk) will provide a dataset on tardigrades and measured microbiomes with accompanying phylogeny.

References

[1] Q. Hong, G. Chen, and Z.-Z. Tang. A phylogeny-based test of mediation effect in microbiome. *ArXiv e–prints*, 2021.

[2] Q. Hong, G. Chen, and Z.-Z. Tang. PhyloMed: a phylogeny-based test of mediation effect in microbiome. *Genome Biology*, 24:24, 2023.

[3] A. Kaul, O. Davidov, and S. D. Peddada. Structural zeroes in high–dimensional data with applications to microbiome studies. *Biostatistics*, 18(3):422–433, 2017.

[4] A. Kaul, S. Mandal, O. Davidov, and S. D. Peddada. Analysis of microbiome data in the presence of excess zeroes. *Frontiers in Microbiology*, 8:2114, 2017.

[5] J. Li. Classification of microbiome data with structural zeroes and small samples. Master's thesis, Linköping University, 2021.

[6] M. Mioduchowska, K. Zając, K. Bartoszek, P. Madanecki, J. Kur, and T. Zając. 16S rRNA gene–based metagenomic analysis of the gut microbial community associated with the dui species *Unio crassus* (Bivalvia: Unionidae). *J. Zool. Syst. Evol. Res.*, 58:615–623, 2020.