Applying data mining tools to infer species community structures from omics data

Agostinetto G.[1]*, Di Filippo M.[2], Sandionigi A.[1], Pescini D.[2], Casiraghi M.[1]
*g.agostinetto@campus.unimib.it
[1] Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan (Italy)
[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan (Italy)

Keywords: metagenomics, high-throughput sequencing (HTS), co-occurrence, networks, data mining

Omics data and HTS sequencing have changed the way to study biology and biodiversity in several fields. In particular, technology advancement enables us to determine taxa composition in very different environments [1], but also to study complex interactions in species communities, as for the human microbiome [2]. Several studies revealed the complexity to infer species associations, highlighting biases linked to sampling strategies or technical issues. Usually, it is challenging to determine relationships between species that cannot be directly studied, as for non-cultivable bacteria, and distinguish the real interaction from spurious information. Our work aim is to characterize community structures with a view to multilevel co-occurrence networks, starting from molecular data generated by HTS. Relying on public repositories, we exploit data mining techniques to reconstruct patterns from metagenomic data [3]. In order to build useful methods to determine biological entities' interactions, we decided to test the efficacy of a commonly used strategy in pattern detection, the Apriori algorithm [4]. We propose a benchmarking approach based on well-characterized community of bacteria derived from HTS experiments, derived from human microbiome projects. Our work is focused not only on the detection of biological rules but also on associations that derived from technical aspects, hiding informative interactions. We think that this work could help to examine how to distinguish real biological interactions from spurious associations of known community, revealing functionality aspects, with the idea to integrate the method in community studies where species dynamics are still a riddle.

References

[1] Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high−throughput genomic approaches for biodiversity analysis. Molecular ecology, 27(2), 313-338.
[2] Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., & Giglio, M. G. (2012). Structure, function and diversity of the healthy human microbiome. nature, 486(7402), 207.
[3] Naulaerts, S., Meysman, P., Bittremieux, W., Vu, T. N., Vanden Berghe, W., Goethals, B., & Laukens, K. (2013). A primer to frequent itemset mining for bioinformatics. Briefings in bioinformatics, 16(2), 216-231.
[4] Tandon, D., Haque, M. M., & Mande, S. S. (2016). Inferring intra-community microbial interaction patterns from metagenomic datasets using associative rule mining techniques. PloS one, 11(4), e0154493.