VIEW IT DIFFERENTLY: FINDING GROUPS IN MICROBIOME DATA

Laura Anderlucci [1], Silvia Dallari[1] and Angela Montanari[1]

[1] Department of Statistical Sciences, University of Bologna, (e-mail: laura.anderlucci@unibo.it, silvia.dallari2@unibo.it, angela.montanari@unibo.it)

ABSTRACT: Microbiota plays a crucial role in human health. Recently, NGS technologies have enabled the exploration of the microbiome without isolation and culturing. However, analyzing and translating microbiome data into meaningful biological insights is challenging due to the data's compositional nature, high dimensionality, sparseness, and over-dispersion. The gut microbiome can vary from individual to individual, and microbiome communities can be grouped to identify community types linked to environmental or health conditions. Different data features, such as individual profiles, community-based descriptors, or genera interactions within a community, provide different perspectives on microbiome complexity. Combining these perspectives could lead to a more comprehensive understanding of microbiome data.

KEYWORDS: model-based clustering, community diversity measures, network-based clustering, consensus clustering

1 Introduction

Microbiota is largely recognized as being a central player in the human health and in that of all organisms and ecosystems, and subsequently has been the subject of intense study. Recently, Next Generation Sequencing (NGS) technologies have enabled the exploration of microbiome without the need for isolation and culturing. The data we are going to study have been obtained through deep sequencing of 16SrRNA genes and grouping bacteria at a certain level of 16SrRNA gene similarity. The analysis and the translation of microbiome data into meaningful biological insights remain still very challenging, also due to particular data characteristics. Microbiome data, in fact, are taxa counts that are compositional in nature (Gloor *et al.*, 2017), high-dimensional, sparse and over-dispersed. In humans, gut microbiome can vary from individual to individual and individual microbiome communities can be grouped to identify community types whose variability can be differently linked to environmental or health conditions.

According to the literature on microbiome data (Xia *et al.*, 2018), different data features can provide different perspectives on microbiome complexity.

The focus has typically been placed either on individual profiles or on community-based descriptors or on genera interactions within a community. We argue that combining these different perspectives could provide a more comprehensive understanding.

2 Microbiome data views

2.1 Individual profiles

The basic sampling units, over which conclusions are generalized, are biological samples. It is of interest to highlight similarities and differences across these units. The fundamental features with which to describe samples are the counts of bacterial species. For interpretation, it is common to imagine prototypical units which can be used as a point of reference for observed samples. In microbiome analysis, these are called *communities*: different communities have different bacterial signatures.

It is worth noticing that this kind of data structure closely resembles the term-document matrix, typically used in the analysis of textual data, and that microbiome data share many of its pros and cons (Sankaran & Holmes, 2019).

2.2 Diversity measures

Characteristic of biological communities is the *biodiversity*, and it can be described either focusing on within-individual richness of taxa or on inter- individual variability. α-diversity is the diversity within a single sample and can be measured via Shannon-Wiener diversity index H' or via Simpson diversity index D:

$$H' = -\sum_{i=1}^{p} p_i \log p_i, \qquad D = 1 - \sum_{i=1}^{p} p_i^2$$

where p_i is the proportion of individuals (or relative abundance) of species i in the community and p is the total number of species present.

β-diversity evaluates differences between two or more units or local assemblages, thus allowing to describe how many taxa are shared between communities or individuals. Examples are the Bray-Curtis dissimilarity:

$$BC = \frac{\sum_{i=1}^{p} |X_{ij} - X_{ik}|}{\sum_{i=1}^{p} (X_{ij} + X_{ik})}$$

where X_{ij}, X_{ik} are the number of individuals in species i in each sample (j,k) and p is the total number of species in samples, and the UniFrac distance. The unweighted (d^U) and weighted (d^W) UniFrac distances exploit the phylogenetic tree information and can be found for two communities A and B as

$$d^U = \sum_{t=1}^{T} \frac{b_t|I(p_t^A > 0) - I(p_t^B > 0)|}{\sum_{t=1}^{T} b_t}, \qquad d^W = \frac{\sum_{t=1}^{T} b_t|p_t^A - p_t^B|}{\sum_{t=1}^{T} b_t(p_t^A + p_t^B)}$$

where p_t^A and p_t^B are the taxa proportions descending from the branch t for community A and B, respectively, T is the rooted phylogenetic tree's branches and b_t is the length of the branch t.

2.3 Network structures

The interactions among the constituent members of a microbial community play a major role in determining the overall behavior of the community and the abundance levels of its members (Xia *et al.*, 2018). These interactions can be modeled using a network whose nodes represent microbial taxa and edges represent pairwise interactions. It is often unreasonable to expect that a single network is able to account for all the interactions in a community and network clustering can help in detecting microbiome features connected, for instance, with different health and environment condition.

3 Microbiome multi-view clustering

Clustering individual profiles (view 1) can be performed via partitioning and hierarchical methods (such as, e.g., spherical k-means, Partitioning Around Medoids, Ward's method) or via model-based methods such as mixtures of Von Mises-Fisher distributions, Dirichlet Multinomial Mixtures, Latent Dirichlet Allocation (see, for a review, Sankaran & Holmes, 2019).

In view of the analogy between microbiome and textual data, we propose to use here the method proposed in Anderlucci *et al.*, 2019, which models the clustering structure through a cosine distance-based mixture. Specifically, given the cosine dissimilarity $d(\mathbf{x}, \xi)$ of a generic sample/document $\mathbf{x}$ from a centroid, say ξ, a distance-based density can be constructed as:

$$f(\mathbf{x}; \xi, \lambda) = \psi(\lambda) e^{-\lambda d(\mathbf{x}, \xi)}$$

where λ is a positive precision parameter and $\psi(\lambda)$ is a normalization constant. In order to perform clustering, we consider a mixture of K cosine distance-based density functions:

$$f(\mathbf{x};\xi,\lambda) = \sum_{k=1}^{K} \pi_k \psi(\lambda) e^{-\lambda d(\mathbf{x},\xi_k)}$$

with positive mixture weights π_k, summing to unity and component varying centroid vectors ξ_k.

When the focus is on community diversity (view 2), the different diversity measures can be combined in a Gower's-coefficient-like fashion in order to guide the clustering of the individuals.

Finally, when the aim is to capture the interaction structure between taxa (view 3) network-based clustering via mixtures of Multivariate Poisson Log-Normal distributions can be applied (Tavakoli & Yooseph, 2019).

The clustering results of the three data views will be combined via consensus clustering (Hornik, 2005) or via the Bayesian two-way latent structure model proposed in Swanson *et al.*, 2019. The proposed multi-view clustering method will be applied to real data on gut microbiome described in McDonald *et al.*, 2018.

References

ANDERLUCCI, L., MONTANARI, A., & VIROLI, C. 2019. The importance of being clustered: Uncluttering the trends of statistics from 1970 to 2015. *Statistical Science*, **34**, 280–300.

GLOOR, G.B., MACKLAIM, J.M., PAWLOWSKY-GLAHN, V., & EGOZCUE, J.J. 2017. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.

HORNIK, K. 2005. A CLUE for CLUster ensembles. *J. Stat. Softw.*, **14**, 1–25.

MCDONALD, D., HYDE, E., *et al.* 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, **3**(3), e00031–18.

SANKARAN, K., & HOLMES, S.P. 2019. Latent variable modeling for the microbiome. *Biostatistics*, **20**(4), 599–614.

SWANSON, D.M., LIEN, T., BERGHOLTZ, H., SØRLIE, T., & FRIGESSI, A. 2019. A Bayesian two-way latent structure model for genomic data integration reveals few pan-genomic cluster subtypes in a breast cancer cohort. *Bioinformatics*, **35**(23), 4886–4897.

TAVAKOLI, S., & YOOSEPH, S. 2019. Learning a mixture of microbial networks using minorization–maximization. *Bioinformatics*, **35**(14).

XIA, Y., SUN, J., & CHEN, D.G. 2018. *Statistical analysis of microbiome data with R*. Singapore: Springer.