Microbiome compositional data analysis for survival studies

Meritxell Pujolassos[1], *Antoni Susín*[2], *M.Luz Calle*[3]

[1]meritxell.pujolassos@uvic.cat, Bioscience Department, Faculty of Sciences, Technology and Engineering, University of Vic – Central University of Catalunya, Vic, Spain

[2] toni.susin@upc.edu, Mathematical Department, UPC-Barcelona Tech, Barcelona, Spain

[3]malu.calle@uvic.cat, Bioscience Department, Faculty of Sciences, Technology and Engineering, University of Vic – Central University of Catalunya, Vic, Spain

Abstract

The compositional nature of microbiome data requires specific compositional data analysis (CoDA) methods. We present a new methodology for the identification of microbial signatures in time-to-event studies. The algorithm implements a CoDA adaptation of elastic-net penalized Cox regression and is integrated in the R package *coda4microbiome* as an extension of the existing functions for cross-sectional and longitudinal studies.

Keywords: microbiome, compositional data analysis, survival data

1. Introduction

Human microbiome is the complete set of microbes found in our bodies, and it plays an important role on human health. Challenging experimental and computational analysis are required to investigate the presence of different microorganisms and understand the complex interactions between them and the environment. High throughput sequencing techniques (16S rRNA and shotgun sequencing), used for identification and quantification of microbial communities, have a limited sequencing capacity which limits the total number of reads that can be revealed from the sample. This total sum restriction implies a great dependence between bacterial species in the analysed sample [1]. Data constraint to a total sum is called compositional data. Compositions are vectors of real positive numbers that contain relative information, which means that each part of the composition on its own is not informative. Information of a composition is extracted from the relation between two or more components [2]. Microbiome data is compositional, therefore, performing its analysis with methods that do not take in account its compositionality may lead to inaccurate results [3], [4].

Compositional Data Analysis (CoDA) was stablished by Aitchison in 1982 [5], who introduced the so-called log-ratio approach, that consists of analysing logarithms of ratios between components instead of each component separately.

coda4microbiome [6] is an algorithm for microbiome analysis based on the log-ratio approach that aims to find a model (microbial signature) with the highest prediction accuracy. *coda4microbiome* has been implemented as an R package and it is developed for cross-sectional and longitudinal studies. In this work, we present a new methodology that extends *coda4microbiome* algorithm to survival data.

The new approach implements elastic-net penalized Cox regression conveniently adapted to CoDA to identify a set of microbial species, and maybe other variables, associated to survival time, i.e., the time until the occurrence of an event of interest, such as, disease onset, response to a treatment, remission, or death.

2. Methods

coda4microbiome algorithm is developed to characterize a microbial signature that best predicts the response variable, and it is structured in three main steps: modelling, variable selection and reparameterization. (1) A regression model with all pairwise log-ratios of microbial species is considered (modelling step), followed by (2) a variable selection step with elastic-net penalization that identifies those log-ratios more associated to the outcome; finally (3) the linear predictor of the log-ratio model is reparametrized to obtain a microbial signature written in terms of the selected bacteria, instead of pairs of bacteria (reparameterization step). Bellow we describe the new *coda4microbiome* algorithm for survival studies.

Assume a survival study with *n* subjects where the time when the event of interest occurs for subject *i* is denoted as t_i. Let $X_i = (X_{i1}, X_{i2}, \dots, X_{iK})$ be the microbial composition for *K* taxa in the *i*-th subject. Microbial abundances (*X*) can be either raw counts or relative abundances. The goal of this method is to identify those microbial taxa whose relative abundances are associated to survival time.

We consider the Cox's proportional hazard regression model (1970) [7] with all possible pairwise log-ratios of taxa as covariates (1). This regression model finds the relationship between pairs of microbes (log-ratios) and the risk of the given event to occur. Using log-ratios in the model, CoDA's principal of scale invariance is ensured.

$$h(t|x) = h_0(t) \cdot \exp\left(\sum_{1\leq j<k\leq K} \beta_{jk} \cdot \log\left(X_j/X_k\right)\right) \quad (1)$$

Variable selection is carried out by the estimation of the regression coefficients (β_{jk}) subjected to an elastic-net penalization (2) where *L* is the loss function for (1). This step allows the removal of those log-ratios less associated to the outcome, thus only log-ratios with non-zero coefficients are kept. Such penalization can also be written in terms of λ and α, which control the amount of penalization and the mixing between norms, respectively (3).

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{L(\beta) + \lambda_1\|\beta\|_2^2 + \lambda_2\|\beta\|_1\} \quad (2)$$

$$\lambda_1 = \lambda(1-\alpha);\ \lambda_2 = \lambda\alpha \quad (3)$$

By default, α is set to 0.9 (adjustable by the user), and optimal λ value is selected after a cross-validation process performed by *cv.glmnet()* from glmnet R package [8]. *coda4microbiome* also allows non-compositional variables (i.e., age, sex, clinical variables, etc) to adjust the model.

After modelling and variable selection, the result is a Cox model composed by the logarithms of pairs of bacteria with non-zero coefficient that are associated to the outcome. The linear term of the model, the linear predictor (right side of equation 4), provides an individual prediction (microbial signature score) for the survival time.

$$M_i = \log\left(h_i(t)/h_0(t)\right) = \sum_{1\leq j<k\leq K} \hat{\beta}_{jk} \cdot \log\left(X_j/X_k\right),\ i\in\{1,\dots,n\} \quad (4)$$

The linearity of logarithms permits the reparameterization of (4) into single bacterial species, instead of pairs of bacteria, which makes interpretation of results more meaningful.

$$M_i = \sum_{1\leq j\leq K} \hat{\theta}_j \cdot \log\left(X_j\right) \quad (5)$$

This final microbial signature is a log-contrast function, *i.e.* $\sum_{j=1}^{K} \hat{\theta} = 0$, which ensures the scale invariant CoDA principle. It also provides a convenient interpretation of the signature as a weighted

balance between two groups of bacteria, those with a positive coefficient vs those with a negative coefficient [9].

3. Results

To exemplify the application of the proposed methodology we used a dataset of simulated intestinal microbiome data from non-obese diabetic mice proposed by Koh, 2018 [10]. Samples from control mice at six weeks of age which were not exposed to any antibiotic treatment were used as template. Survival time, event (developing/not developing diabetes), censoring, age, and sex were also simulated in Koh's dataset. A final dataset of 100 samples and 353 different bacteria was used to perform the survival analysis using *coda4microbiome*. The aim of the analysis was to characterize a microbial signature able to predict the risk of developing diabetes.

The initial Cox model performed in the modelling step contained all possible pairwise log-ratios from the 353 bacteria's relative abundances, and it was adjusted by age and sex. Variable selection was performed with elastic-net and the optimal penalized parameter was stablished by cross-validation. It resulted in a model of 4 pairs of log-ratios with a mean cross-validation Harrell C index of 0.64 (±0.05). After reparameterization, the final microbial signature was expressed as a weighted balance between 4 bacterial species with positive coefficient and 3 with negative coefficient (Figure 1).

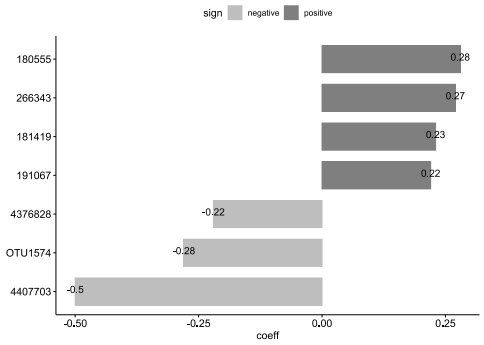

Figure 1: Bacterial species (vertical axis) with their respective coefficients that compose the microbial signature for the survival data analysis with *coda4microbiome*.

4. Conclusions

We introduced a new methodology for microbial analysis in survival studies that accounts for the compositional nature of microbiome data. The algorithm identifies a microbial signature that predicts the risk of a given event with the highest accuracy. Such signature is expressed as a balance between two groups of bacteria.

The algorithm has been implemented in R and it has been integrated in the existing R package *coda4microbiome* so that it can be easily used in survival studies to identify which microbial species are more associated to the development of a disease, response to a treatment, or even death.

5. Acknowledgments

This work was partially supported by the Spanish Ministry of Economy, Industry and Competitiveness, references PID2019-104830RB-I00 (M.L.C), PID2021-123657OB-C33 (A.S) and PID2021-122136OB-C21 (A.S.).

6. Bibliography

[1] Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics and Informatics*, *17*(1), e6. https://doi.org/10.5808/GI.2019.17.1.e6

[2] Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, *70*(1), 57–65.

[3] Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 1–18. https://doi.org/10.1186/s40168-017-0237-y

[4] Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M. A., Wright, R. J., Dhanani, A. S., Comeau, A. M., & Langille, M. G. I. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, *13*(1), 1–16. https://doi.org/10.1038/s41467-022-28034-z

[5] Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society*, *44*(2), 139–177.

[6] Calle, M. L., & Susin, A. (2022). coda4microbiome: compositional data analysis for microbiome studies. *BioRxiv*, 2022.06.09.495511. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.06.09.495511v1

[7] Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220. http://www.jstor.org/stable/2985181

[8] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/JSS.V033.I01

[9] Susin, A., Wang, Y., Cao, K. A. L., & Luz Calle, M. (2020). Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, *2*(2). https://doi.org/10.1093/nargab/lqaa029

[10] Koh, H., Livanos, A. E., Blaser, M. J., & Li, H. (2018). A highly adaptive microbiome-based association test for survival traits. *BMC Genomics*, *19*(1), 1–13. https://doi.org/10.1186/s12864-018-4599-8