Bioinformatics, 36(7), 2020, 2298–2299
doi: 10.1093/bioinformatics/btz862
Advance Access Publication Date: 18 November 2019
Applications Note

Systems biology

Link-HD: a versatile framework to explore and integrate heterogeneous microbial communities

Laura M. Zingaretti [1,2,*], Gilles Renand[3], Diego P. Morgavi[4] and Yuliaxis Ramayo-Caldas[3,5]

[1]Plant and Animal Genomics, Statistical and Population Genomics Group, CSIC-IRTA-UAB-UB Consortium, Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain, [2]IAPCBA and IAPCH, UNVM, Villa María, Córdoba 5900, Argentina, [3]URM Animal Genetics and Integrative Biology, GABI, INRA, AgroParisTech, Université Paris-Saclay, 78352 Jouy-en-Josas, France, [4]Animal Physiology and Livestock Systems Divisions, INRA, Herbivore Research Unit, Clermont Auvergne University, Saint Genès-Champanelle 63122, France and [5]Animal Breeding and Genetics Program, IRTA, 08140 Caldes de Montbui, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 19, 2019; revised on October 4, 2019; editorial decision on November 14, 2019; accepted on November 15, 2019

Abstract

Motivation: We present Link-HD, an approach to integrate multiple datasets. Link-HD is a generalization of 'Structuration des Tableaux A Trois Indices de la Statistique–Analyse Conjointe de Tableaux', a family of methods designed to integrate information from heterogeneous data. Here, we extend the classical approach to deal with broader datasets (e.g. compositional data), methods for variable selection and taxon-set enrichment analysis.

Results: The methodology is demonstrated by integrating rumen microbial communities from cows for which methane yield (CH_4y) was individually measured. Our approach reproduces the significant link between rumen microbiota structure and CH_4 emission. When analyzing the TARA's ocean data, Link-HD replicates published results, highlighting the relevance of temperature with members of phyla Proteobacteria on the structure and functionality of this ecosystem.

Contact: m.lau.zingaretti@gmail.com

Availability and implementation: The source code, examples and a complete manual are freely available in GitHub https://github.com/lauzingaretti/LinkHD and in Bioconductor https://bioconductor.org/packages/release/bioc/html/LinkHD.html.

1 Introduction

The reduction of 'omics' technology costs now enables collection of data from multiple sources. This allows researchers to simultaneously study several datasets and investigate their relationship with complex traits. The integration of these heterogeneous datasets is not trivial and several statistical methods have been developed to address this challenge (Argelaguet *et al.*, 2018; Mariette and Villa-Vialaneix, 2018; Meng *et al.*, 2014). In particular, the amalgamation of multiple microbial ecosystems poses unique challenges as these are compositional and sparse data. MixKernel (Mariette and Villa-Vialaneix, 2018) is a well-known tool designed to integrate heterogeneous datasets including microbial communities, but no method to perform a taxonomic enrichment analysis is available. Another popular integrative approach is MOFA (Argelaguet *et al.*, 2018), however, it is unable to deal with compositional data.

Here, we present Link-HD, a tool to integrate and explore multiple microbial communities based on STATIS (Des Plantes, 1976), a family of multivariate methods to integrate multiple datasets. Link-HD generalizes STATIS with Regression Biplot (Ter Braak, 1997), clustering, differential abundance, enrichment taxonomic analysis and visualization tools. Link-HD analyzes distance tables computed from numerical, categorical, or compositional data as a generalization of multidimensional scaling (Abdi *et al.*, 2007). Furthermore, Link-HD performs variable selection and can link the obtained common sub-space with phenotype information.

2 Materials and methods

Like STATIS, Link-HD aims to compare and analyze the relationships between datasets with a shared set of observations or variables. However, our package was specifically designed to integrate microbial communities and incorporate distances and transformations to deal with compositional data (Aitchison, 1982). The method is implemented in three main phases (Fig. 1).

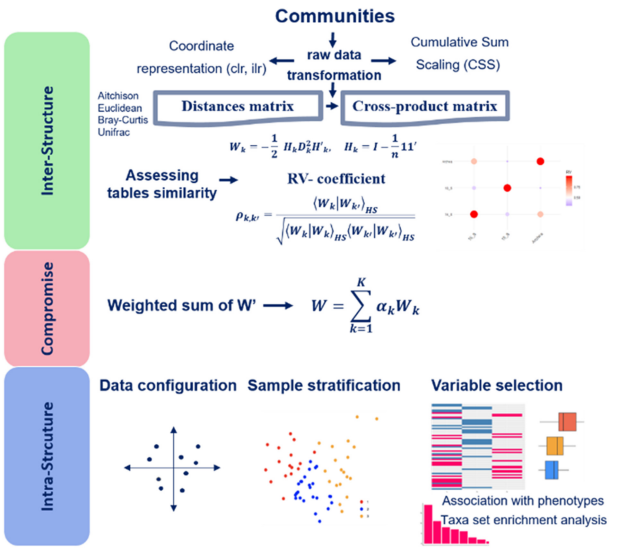

Fig. 1. Link-HD Workflow. In the Inter-structure step, raw data are transformed using cumulative sum scaling or centered log ratio, and the correlation coefficient (Rv) is computed. The second step is the compromise (W) and, finally, the intra-structure step involves the Eigen-decomposition of W. Observations can be clustered and methods for selecting variables and association with phenotypes are available

i. Inter-structure step: The algorithm first assesses the similarity between transformed distance tables using the vector correlation coefficient (Rv) (Escoufier, 1973), which can be interpreted as a general 'vector covariance' between matrices, i.e. this step evaluates similarity between the disparate datasets.
ii. Compromise step: Next, the 'compromise' matrix is calculated, which is a weighted sum of each cross-product matrix. This step involves an optimization problem since the weights are chosen to maximize the correlation between the compromise matrix and each individual component.
iii. Intra-structure step: Finally, the compromise matrix is evaluated through a Principal Component Analysis. The coordinates of the common elements are projected into a low rank space, where the relationships between them can be easily interpreted.

Variable selection is tackled by two alternative approaches: (i) by projecting all the input variables into the compromise through a general Biplot formulation (Ter Braak, 1997); and (ii) by computing the differential abundance of features between clusters of samples. A novelty of Link-HD is its ability to aggregate the selected variables at several taxonomic levels and to establish whether that level is enriched using a cumulative hypergeometric distribution. This function also allows users to add a custom OTUs list. Finally, the SPIEC-EASI (Kurtz *et al.*, 2015) tool can be used to visualize variable interactions.

3 Case studies

We illustrate our approach with rumen microbial (Ramayo-Caldas *et al.*, 2019), TARA's Ocean expedition (Sunagawa *et al.*, 2015) and transcriptome NCI-60 cell line datasets (Reinhold *et al.*, 2012).

In the rumen study, we integrated Bacteria, Archaea and Protozoa from 65 Holstein cows. Link-HD was able to reproduce previous results (Danielsson *et al.*, 2017; Kittelmann *et al.*, 2014; Ramayo-Caldas *et al.*, 2019), showing a link between the structure of the rumen microbiota and CH_4 emission. We also identify microbial markers associated to CH_4. In the TARA's example, Link-HD replicates the relevant role of temperature and Proteobacteria phyla on the structure of this ecosystem, as described in Mariette and Villa-Vialaneix (2018). Finally, we show the potential of Link-HD to integrate other omics layers by using transcriptome NCI-60 cell lines. Link-HD recapitulates the reported data structure (Meng *et al.*, 2014) and ontology analysis reveals several cancer-related pathways.

In all, our results demonstrate that Link-HD is robust in combining several heterogeneous data types. A detailed description of these case studies and the theory behind Link-HD is available at https://lauzingaretti.github.io/LinkHD/ and in Bioconductor (https://bioconductor.org/packages/release/bioc/html/LinkHD.html).

4 Conclusions

We have developed an R package to integrate multiple microbial communities and other 'omics' layers combining a plethora of statistical methods in a fast, simple and flexible way.

Acknowledgements

The authors would like to thank Miguel Pérez-Enciso, Rayner González-Prendes and Jordi Estelle for valuable discussions and comments on the manuscript and also to APIS-GENE for funding the MicroFicient project.

Funding

L.M.Z. is recipient of a Ph.D. grant from Ministry of Economy and Science, Spain associated with 'Centro de Excelencia Severo Ochoa 2016–2019' award SEV-2015-0533 to CRAG. Y.R.C. was funded by Marie Skłodowska-Curie grant (P-Sphere) agreement No 6655919 (EU).

Conflict of Interest: none declared.

References

Abdi,H. *et al.* (2007) Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Qual. Prefer.*, **18**, 627–640.
Aitchison,J. (1982) The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B*, **44**, 139–160.
Argelaguet,R. *et al.* (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, **14**, e8124.
Danielsson,R. *et al.* (2017) Methane production in dairy cows correlates with rumen methanogenic and bacterial community structure. *Front. Microbiol.*, **8**, 226.
Des Plantes,H.L. (1976) Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe. Doctoral dissertation, Université des sciences et techniques du Languedoc.
Escoufier,Y. (1973) Le traitement des variables vectorielles. *Biometrics*, **29**, 751–760.
Kittelmann,S. *et al.* (2014) Two different bacterial community types are linked with the low-methane emission trait in sheep. *PLoS One*, **9**, e103171.
Kurtz,Z.D. *et al.* (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.*, **11**, e1004226.
Mariette,J. and Villa-Vialaneix,N. (2018) Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, **34**, 1009–1015.
Meng,C. *et al.* (2014) A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, **15**, 162.
Ramayo-Caldas,Y. *et al.* (2019) Identification of rumen microbial biomarkers linked to methane emission in Holstein dairy cows. *J. Anim. Breed. Genet.*, doi: 10.1111/jbg.12427.
Reinhold,W.C. *et al.* (2012) CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.*, **72**, 3499–3511.
Sunagawa,S. *et al.* (2015) Ocean plankton. Structure and function of the global ocean microbiome. *Science*, **348**, 1261359.
Ter Braak,C.J.F. (1997) JC Gower and DJ Hand, biplots. Monographs on statistics and applied probability. *Psychometrika*, **62**, 457–460.