NeighborFinder: an R package inferring local microbial network around a species of interest

Mathilde Sola[1,2], Adrien Paravel[3], Sandrine Auger[3], Jean-Marc Chatel[3], Florian Plaza-Oñate[1], Emmanuelle Le Chatelier[1], Marion Leclerc[1,5], Patrick Veiga[1,3], Clémence Frioux[2], Mahendra Mariadassou[4,*], Magali Berland[1,*]

[1]Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France

[2]Inria, University of Bordeaux, INRAE, 33400, Talence, France

[3]Université Paris-Saclay, INRAE, MICALIS, 78350, Jouy-en-Josas, France

[4]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

[5]Université Clermont Auvergne, INRAE, UMR454 MEDIS, 63000, Clermont-Ferrand, France

* Corresponding authors

1 Abstract

Motivation: Understanding interactions from microbiome data is a central aspect in microbial ecology, as it provides insights into ecosystem stability, disease mechanisms, and can be used to design synthetic communities. Current network inference tools reconstruct global networks from co-abundance data, which means they capture the overall correlation structure for the entire set of taxa considered. These approaches are computationally intensive and suboptimal when the focus is on the local neighborhood of specific taxa of interest.

Results: We introduce NeighborFinder, a local network inference method that enables the targeted discovery of direct neighbors around a species of interest. Using cross-validated multiple linear regression with ℓ_1 penalty and microbiome-specific filters, our approach infers interpretable species-centered interactions, with F1 score ≥ 0.95 on simulated cohorts ranging from 250 to 1000 samples. This method is well-suited for large metagenomic datasets and is particularly valuable for exploratory studies where the targeted hypotheses outweigh the need for global community structure. The approach complements existing methods by being a biologically intuitive and computationally efficient.

Availability and Implementation: The R package is freely available on GitHub: `https://github.com/metagenopolis/NeighborFinder`. The data and source code used to calculate performances and produce the use case example in this paper can be found respectively at: `https://doi.org/10.57745/UPITJ0` and `https://doi.org/10.57745/HJLWW4`.

Supplementary information: Supplementary data are available

1.1 Keywords

metagenomics, microbiome, network inference, local network, companion microbial species

2 Introduction

In recent years, metagenomic sequencing has revolutionized our understanding of microbial communities and enabled researchers to identify species, estimate their abundances, and infer functional potential. As metagenomic datasets grow in scale and resolution, so does the interest in moving towards a more mechanistic understanding of microbial communities. Microbial networks, where the nodes are taxa and the edges represent co-abundances - that can be interpreted as potential ecological associations - offer a powerful framework to explore the structure of these communities [1].

Researchers may be primarily interested in identifying the species that interact directly with a specific species of interest (*e.g.* a pathogen, a probiotic) rather than reconstructing the full network [2]. Identified neighbors could for instance support the design of consortia promoting the growth of a species [3].

Existing network inference methods, many of which are based on Graphic Gaussian Models (GGM) [4], typically rely on global modeling approaches and require inference over the full set of taxa. However, they do not directly address the question of identifying the local neighbors of a specific species of interest.

Here, we introduce NeighborFinder, a local inference method integrated into an R package that enables fast (running time <1 minute), efficient and scalable identification of direct neighbors around a species of interest.

3 Overview of the method

NeighborFinder is tailored to microbiome data. It was specifically developed for shotgun metagenomic data and includes a default normalization step for such datasets, but can accommodate metabarcoding data (and other count-based inputs) by skipping this step.

The primary function for inferring the local neighborhood of a species is `apply_neighborfinder()`. The R package provides helper functions to visualize results, either as a network or in tabular form, and to construct outputs based on the intersection of multiple inferences. A technical report and a vignette are available to illustrate typical use cases and provide methodological details.

In this application note, we show how NeighborFinder infers local networks from bacterial abundances in shotgun metagenomic data, focusing on the neighborhoods of three well-known species of interest.

3.1 NeighborFinder description

NeighborFinder performs a three-step procedure illustrated in Fig. 1A and detailed below.

1) Data preparation. Input data is a species by sample abundance table. The first step is to filter the table using the parameter `prev_level`. This filtering simplifies the dataset by filtering out low-prevalence species, for which network interaction detection is notoriously difficult, especially in studies with limited sample sizes [5]. Next, shotgun metagenomic abundances are transformed into count-like values using `get_count_table()`, by dividing all values by the smallest positive abundance and rounding the results, so that the minimum non-zero count is 1. Finally, `mclr` normalization [6] is applied to the previously transformed abundance data, an adapted version of the clr normalization that preserves the zeros in the dataset.

2) Network inference. This step considers a linear regression problem where the abundance of the species of interest is regressed against the abundances of all other species. The function `glmnet::cv.glmnet()` [7] is applied to the normalized data 10 times, each run using a different Random Number Generator (RNG) seed. This function incorporates a ℓ_1 penalization term λ enforcing the strength of the regularization and thus the number of non-null coefficients. This is solved using `glmnet::glasso()` and cross-validation is used to tune the parameter λ. For each RNG seed, a filtering step is further applied to increase the reliability of detected interactions. This is done with the `top_filtering` parameter (set to 30 by default), which keeps only the topmost coefficients in absolute value (*e.g.* the top 30% if `top_filtering = 30`). This filter is applied to limit spurious detection of neighbors.

3) Network stabilization. To improve robustness, only neighbors consistently detected across multiple runs are kept. This step merges the 10 filtered outputs (one per RNG seed) and discards edges detected in less than half of them. For the remaining edges, the final coefficient is computed as the median over all inferences, and they are gathered in an edge table.

3.2 Performance of the method

3.2.1 Simulation strategy

The function `apply_neighborfinder()` has two key parameters (`prev_level` and `top_filtering` shown in Fig. 1A) that shape the output. It is also acknowledged that the dataset size usually has an impact on the generated network [4].

Our performance assessment strategy therefore consisted in covering a range of values for both parameters, with the aim of providing users with guidelines to pick the most suitable combination depending on their dataset size.

We tested the performance of NeighborFinder on eight shotgun metagenomic cohorts (sample sizes ranging from 347 to 1084). For each cohort, similar microbiome data was simulated with known interaction networks (see Supplementary Material). NeighborFinder's performance was evaluated using precision and recall, aggregated into the F1 score (harmonic mean of recall and precision).

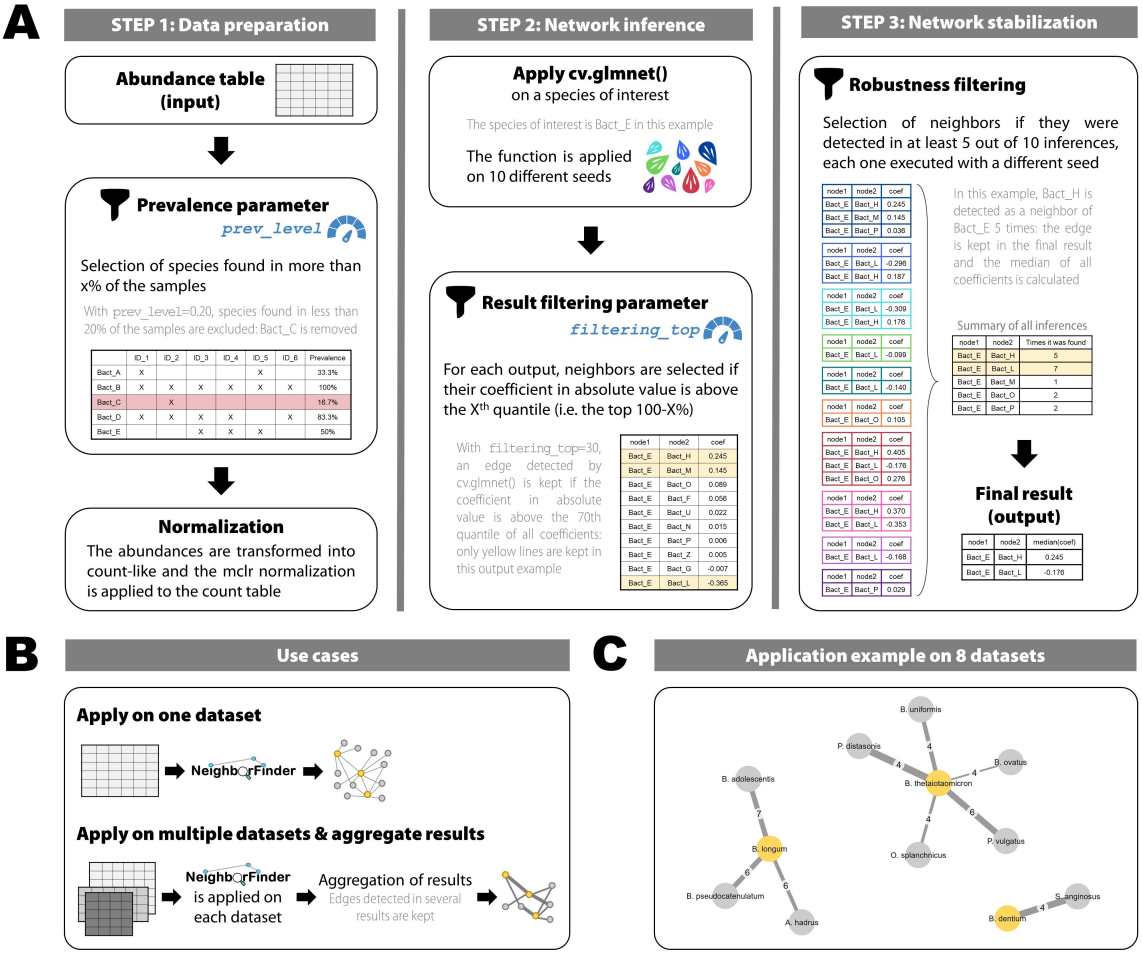

Figure 1: **Overview of NeighborFinder, its use cases and an example.** A) Presentation of the key function `apply_neighborfinder()` with metagenomic data as input (microbial abundances for each sample), highlighting the two parameters that can be adjusted by the user: `prev_level` and `filtering_top`. B) Description of two main NeighborFinder use cases: on a single dataset, or on multiple datasets for aggregation purposes. C) Aggregated network of an application example focused on three species of interest (yellow nodes): *Bifidobacterium longum*, *Bifidobacterium dentium*, and *Bacteroides thetaiotaomicron* in eight datasets. The edge label is the number of datasets in which the interaction was detected. The edge width is proportional to the mean coefficient value.

3.2.2 Evaluation procedure

For each bacterial species, local networks were inferred using `prev_level` values between 15% and 35% (in 5% increments). Each setting was repeated 10 times with different RNG seeds, for all simulated datasets. The `glmnet::cv.glmnet()` function was applied, using `cvglm_to_coeffs_by_object()` from the NeighborFinder package. The resulting inferences were then filtered following the procedure implemented in `apply_NeighborFinder()` keeping only the top `filtering_top`% coefficients, ranging from 5% to 30% (in 5% increments), and including 100% to assess unfiltered results (see Fig. 1A, steps 1 and 2).

Finally, for each species, a network was obtained through the stabilization step implemented in NeighborFinder (see Fig. 1A step 3), where edges were retained if detected in at least 5 of the 10 runs across different seeds. F1-scores were calculated for each species, then averaged across species and cohorts.

Table 1: Application example listing detected neighbors of three species of interest, with an aggregation strategy. Abbreviations: coef, coefficient; SD, standard deviation.

Node 1	Node 2	Occurrence	Mean coef	SD	Biological interpretation
Bacteroides thetaiotaomicron	*Bacteroides ovatus*	4/8	0.054	0.016	Functional complementarity [8]
Bacteroides thetaiotaomicron	*Parabacteroides distasonis*	4/8	0.21	0.092	Taxonomic proximity, niche preference [5]
Bacteroides thetaiotaomicron	*Bacteroides uniformis*	4/8	0.13	0.084	Taxonomic proximity, niche preference [5]
Bacteroides thetaiotaomicron	*Odoribacter splanchnicus*	4/8	0.090	0.054	Potential cross-feeding
Bacteroides thetaiotaomicron	*Phocaeicola vulgatus*	6/8	0.18	0.082	Functional complementarity [9]
Bifidobacterium longum	*Anaerostipes hadrus*	6/8	0.11	0.053	Potential cross-feeding [10]
Bifidobacterium longum	*Bifidobacterium pseudocatenulatum*	6/8	0.15	0.040	Taxonomic proximity, niche preference [5]
Bifidobacterium longum	*Bifidobacterium adolescentis*	7/8	0.15	0.058	Taxonomic proximity, niche preference [5]
Bifidobacterium dentium	*Streptococcus anginosus*	4/8	0.20	0.060	Niche preference [11]

3.3 Determining the best combination of parameters

Figure S1 summarizes performance scores calculated on simulated dataset with sizes ranging from n=100 to n=1000, and Figure S2 focuses on n=50 using a wider range of prevalence filters. These results show that the parameters `prev_level` and `filtering_top` should be adjusted according to the size of the dataset. When dealing with large datasets, many parameter combinations led to an equivalent expected performance, with the dark green areas indicating a high F1 score ($\geq$0.9) in Figure S1. One strategy could be to select a low value of `prev_filter` to include as many bacterial species as possible and then choose the `filtering_top` value to ensure satisfactory performance (*e.g.* with n=1000, `prev_level`=0.15 and `filtering_top`=30). As the dataset size decreases, increasing the `prev_level` value while maintaining a high value of `filtering_top` is required to maintain high F1-scores. In particular, for small datasets (n=50), Figure S2 highlights that stringent filtering on prevalence (`prev_level` $\geq$ 0.4) is mandatory to achieve a F1 score of 0.7 or higher.

4 Usage scenario

4.1 Use cases

The method can be applied to one or several datasets (Fig. 1B). In the latter case, we recommend running NeighborFinder independently on each dataset before merging the results. An aggregated network can then be built from edges detected across several datasets (*e.g.* in more than 3 out of 5 datasets). This strategy aims to enhance the overall robustness of the inferred network. An illustration of this strategy is provided in Fig. 1C, with details in Table 1.

4.2 Application example

NeighborFinder was independently applied to each of the eight datasets (see Supplementary Material) to identify neighbors for three species of interest that are well described in the litterature: *Bifidobacterium longum*, *Bifidobacterium dentium*, and *Bacteroides thetaiotaomicron*. After merging results, edges were kept in the aggregated network if they were detected in at least half of the inferences ($\geq$4/8). Fig. 1C shows the resulting network and Table 1 lists for each edge the following descriptors: occurrence (number of times detected), mean coefficient and standard deviation, and biological interpretation of the edge based on literature review.

In particular, the interaction between *B. thetaiotaomicron* and *Bacteroides ovatus* may reflect functional complementarity: *B. ovatus* harbors different glycoside hydrolases and polysaccharide utilization loci (PULs) able to degrade complex plant polysaccharides such as hemicelluloses (galactomannans and xylans) that *B. thetaiotaomicron* lacks [8]. Such complementarity characterizes *B. ovatus* as a companion species to *B. thetaiotaomicron*. Similarly, *B. vulgatus* complements the enzymatic activity of *B. thetaiotaomicron* [9]. The interaction between *B. longum* and *Anaerostipes hadrus* may plausibly involve cross-feeding, as previously demonstrated between *Bifidobacterium* species and butyrate-producing bacteria such as *Anaerostipes* species [10]. Other detected neighbors may be explained by taxonomic

proximity or niche similarity, as often happening with edges detected by network inference methods when the environment is not explicitly accounted for [5]. For example, *B. dentium* and *Streptococcus anginosus* both inhabit the human oral cavity[11].

This application illustrates the variety of biological insights that can emerge from co-abundance patterns in NeighborFinder local networks.

5 Discussion

NeighborFinder provides an efficient and interpretable approach complementary to global network inference. This framework is particularly suited for large-scale datasets and exploratory analyses where speed (see Figure S3 for execution time details), simplicity, and species-centered hypotheses are key.

While it addresses key challenges such as compositionality and sparsity through `mclr` normalization and prevalence filtering to exclude low-prevalence taxa, NeighborFinder suffers from the linearity assumption inherent to GGM-based approaches, which may limit the description of certain types of biological interactions (amensalism, parasitism, etc.) [5]. Similarly to global approaches, our method detects relatively few negative edges. These could be interpreted as signals of competition or inhibition between taxa.

To mitigate instability and noise inherent to all network inference methods, we extended the naive `cv.glmnet` approach with additional features specifically tailored to microbiome data. We used ℓ_1 penalized regression combined with data-specific filtering steps, applied post-processing to the results, and introduced the use of multiple repetitions alongside cross-validation to improve robustness.

Although NeighborFinder has been designed for microbiome data, this method could be useful to explore local networks in other context of tabular data ranging from different taxonomic levels to functional modules for example.

6 Funding

This work was supported by both Carnot Qualiment® (#20 CARN 0026 01) and by the Metagenopolis grant (ANR-11-DPBS-0001).

References

[1] Mehdi Layeghifard, David M. Hwang, and David S. Guttman. Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*, 25(3):217–228, March 2017.

[2] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, June 2006. Publisher: Institute of Mathematical Statistics.

[3] Ravin Poudel, Ari Jumpponen, Megan M. Kennelly, Cary Rivard, Lorena Gomez-Montano, and Karen A. Garrett. Integration of Phenotypes in Microbiome Networks for Designing Synthetic Communities: a Study of Mycobiomes in the Grafted Tomato System. *Applied and Environmental Microbiology*, 89(6):e01843–22, May 2023. Publisher: American Society for Microbiology.

[4] Camille Champion, Raphaëlle Momal, Emmanuelle Le Chatelier, Mathilde Sola, Mahendra Mariadassou, and Magali Berland. OneNet—One network to rule them all: Consensus network inference from microbiome data. *PLOS Computational Biology*, 20(12):e1012627, December 2024. Publisher: Public Library of Science.

[5] Lisa Röttjers and Karoline Faust. From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews*, 42(6):761–780, November 2018.

[6] Grace Yoon, Irina Gaynanova, and Christian L. Müller. Microbial Networks in SPRING - Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data. *Frontiers in Genetics*, 10:516, June 2019.

[7] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[8] Eric C. Martens, Elisabeth C. Lowe, Herbert Chiang, Nicholas A. Pudlo, Meng Wu, Nathan P. McNulty, et al. Recognition and Degradation of Plant Cell Wall Polysaccharides by Two Human Gut Symbionts. *PLOS Biology*, 9(12):e1001221, December 2011. Publisher: Public Library of Science.

[9] Morten Nilsen, Eva Maria Rehbinder, Karin C. Lødrup Carlsen, Guttorm Haugen, Gunilla Hedlin, et al. A Globally Distributed Bacteroides caccae Strain Is the Most Prevalent Mother-Child Shared Bacteroidaceae Strain in a Large Scandinavian Cohort. *Applied and Environmental Microbiology*, 89(7):e00789–23.

[10] Audrey Rivière, Marija Selak, David Lantin, Frédéric Leroy, and Luc De Vuyst. Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut. *Frontiers in Microbiology*, 7:979, June 2016.

[11] Nima Kianoush, Christina J. Adler, Ky-Anh T. Nguyen, Gina V. Browne, Mary Simonian, and Neil Hunter. Bacterial Profile of Dentine Caries and the Impact of pH on Bacterial Population Diversity. *PLOS ONE*, 9(3):e92940, March 2014. Publisher: Public Library of Science.