Open access • Posted Content • DOI:10.1101/2021.07.13.452182

Disentangling environmental effects in microbial association networks

— Source link

Ina M. Deutschmann, Gipsi Lima-Mendez, Anders K. Krabberød, Jeroen Raes ...+3 more authors

Institutions: Spanish National Research Council, Université catholique de Louvain, University of Oslo, Katholieke Universiteit Leuven

Published on: 13 Jul 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Disentangling environmental effects in microbial association networks

Ina Maria Deutschmann[1*], Gipsi Lima-Mendez[2], Anders K. Krabberød[3], Jeroen Raes[4,5], Sergio M. Vallina[6], Karoline Faust[5*†] and Ramiro Logares[1*]

[1] Institute of Marine Sciences, CSIC, Passeig Marítim de la Barceloneta, 37, 08003, Barcelona, Spain.
[2] Louvain Institute of Biomolecular Science and Technology (IBST), Université catholique de Louvain, Croix du sud 4-5/L7.07.02, 1348, Louvain-la-Neuve, Belgium.
[3] Department of Biosciences/Section for Genetics and Evolutionary Biology (EVOGENE), University of Oslo, p.b. 1066 Blindern, N-0316, Oslo, Norway.
[4] VIB Center for Microbiology, Herestraat 49-1028, 3000, Leuven, Belgium.
[5] KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Molecular Bacteriology, Leuven, Belgium, Herestraat 49, 3000, Leuven, Belgium.
[6] Spanish Institute of Oceanography (IEO), Ave Principe de Asturias 70 Bis, 33212, Gijon, Spain.

[*] Corresponding authors:
Ina Maria Deutschmann (ina.m.deutschmann@gmail.com)
Karoline Faust (karoline.faust@kuleuven.be)
Ramiro Logares (ramiro.logares@icm.csic.es)

[†] Shared last authors

Abstract

Background

Ecological interactions among microorganisms are fundamental for ecosystem function, yet they are mostly unknown or poorly understood. High-throughput-omics can indicate microbial interactions through associations across time and space, which can be represented as association networks. Associations could result from either ecological interactions between microorganisms, or from environmental selection, where the associations are environmentally-driven. Therefore, before downstream analysis and interpretation, we need to distinguish the nature of the association, particularly if it is due to environmental selection or not.

Results

We present EnDED (**En**vironmentally-**D**riven **E**dge **D**etection), an implementation of four approaches as well as their combination to predict which links between microorganisms in an association network are environmentally-driven. The four approaches are Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality. We tested EnDED on networks from simulated data of 50 microorganisms. The networks contained on average 50 nodes and 1087 edges, of which 60 were true interactions but 1026 false associations (i.e. environmentally-driven or due to chance). Applying each method individually, we detected a moderate to high number of environmentally-driven edges—87% Sign Pattern and Overlap, 67% Interaction Information, and 44% Data Processing Inequality. Combining these methods in an intersection approach resulted in retaining more interactions, both true and false (32% of environmentally-driven associations). After validation with the simulated datasets, we applied EnDED on a marine microbial network inferred from 10 years of monthly observations of microbial-plankton abundance. The intersection combination predicted that 8.3% of the associations were environmentally-driven, while individual methods predicted 24.8% (Data Processing Inequality), 25.7% (Interaction Information), and up to 84.6% (Sign Pattern as well as Overlap). The fraction of environmentally-driven edges among negative microbial associations in the real network increased rapidly with the number of environmental factors.

Conclusions

To reach accurate hypotheses about ecological interactions, it is important to determine, quantify, and remove environmentally-driven associations in marine microbial association networks. For that, EnDED offers up to four individual methods as well as their combination. However, especially for the intersection combination, we suggest using EnDED with other strategies to reduce the number of false associations and consequently the number of potential interaction hypotheses.

Keywords: microbial interactions; association network; effect of indirect dependencies; environmentally-driven edge detection

Background

Association networks to generate microbial interaction hypotheses

There is a myriad of microorganisms on Earth; current estimates indicate $\approx 10^{12}$ microbial species (Locey & Lennon, 2016), and $\approx 10^{30}$ microbial cells (Whitman *et al.*, 1998; Kallmeyer *et al.*, 2012). Microorganisms have crucial roles in the biosphere by contributing to global biogeochemical cycles (Falkowski *et al.*, 2008) and underpinning diverse food webs. The importance of microbes for the functioning of ecosystems cannot be understood without considering their ecological interactions (DeLong, 2009; Krabberød *et al.*, 2017). These allow transferring carbon and energy to upper trophic levels, and the recycling of nutrients and energy (Worden *et al.*, 2015). Furthermore, ecological interactions influence microbial community turnover and composition. These interactions include win-win (e.g. mutual cross-feeding and cooperation), win-loss (e.g. predator-prey and host-parasite), and loss-loss (e.g. resource competition) relationships (Faust & Raes, 2012). Although microbial communities are highly interconnected (Layeghifard *et al.*, 2017), our knowledge about ecological interactions in the microbial world is still limited (Krabberød *et al.*, 2017; Bjorbækmo *et al.*, 2019).

Previous studies have shown relationships between a restricted number of microorganisms. However, we need a large number of interactions to understand the functioning of complex ecosystems. This is challenging, in part, due to the vast number of possible interactions—given n microorganisms, there are $\binom{n}{2} = n(n-1)/2$ potential pairwise interactions. Thus, it is unfeasible to test them experimentally within a reasonable amount of time and cost. The problem of having a large number of potential interactions can be partially circumvented with omics technologies coupled to network analyses.

Omics can identify and quantify a large number of microorganisms from a given sample. Typically, the relative abundance for each identified organism per sample is estimated. There are multiple methods to determine associations (normally based on correlations) between microorganisms using their abundances (e.g. eLSA (Xia *et al.*, 2011, 2013), CoNet (Faust & Raes, 2016), SPIEC-EASI (Kurtz *et al.*, 2015), or FlashWeave (Tackmann *et al.*, 2019)). These abundance-based associations compose a network, where nodes represent microorganisms and edges represent either co-presence (positive association) or mutual exclusion (negative association) relationships, which constitute microbial interaction hypotheses.

Challenges in using networks as a representation of the microbial ecosystem

Although networks play an essential role in understanding complex systems, microbial ecological networks are not yet as developed in terms of inference and biological interpretation (Lv *et al.*, 2019). Network inference from -omics data is difficult (Li *et al.*, 2016; Layeghifard *et al.*, 2017) because of both technical and interpretation challenges. One challenge is the compositional nature of the data produced by DNA sequencers (Gloor *et al.*, 2017). There are several network tools (Li *et al.*, 2016) that consider this, e.g., SPIEC-EASI (Kurtz *et al.*, 2015). Other difficulties include data based on a small number of samples relative to the number of microorganisms they contain, i.e., a low sample-to-microorganisms ratio; plus sparse data—too many zeros in the dataset that can wrongly associate microorganisms (Aitchison, 1981). A zero indicates either the absence of a microorganism (structural zero), or an insufficient detection level or sequencing depth (sampling zero). Thus, we should remove microorganisms appearing in just a few samples.

Interpretation of association networks is challenging because they are not equivalent to ecological networks. Edges in ecological networks represent observed ecological interactions between different microorganisms like parasitism or competition (Xiao *et al.*, 2017). Ecological networks are directed graphs, where the directed edges (arcs) point from a start node (source) to an end node (target). In contrast, association networks are undirected. Although association networks provide ecological insight, they do not necessarily encode causal relationships or observed ecological interactions. Unless edges are verified with experiments or additional information, one should be careful when attributing biological meaning to network properties (Röttjers & Faust, 2018). In addition, networks with too many edges (dense networks or hairballs) make interpretation more challenging. We can reduce network density when lowering the corrected p-value for inferred edges (Weiss *et al.*, 2016), or increasing the cut-off for other criteria such as the association strength, prevalence, or abundance filtering (Röttjers & Faust, 2018). Another strategy is agglomeration using taxonomic or ecological (functional) groupings (Lima-Mendez *et al.*, 2015).

The interpretation challenge addressed in this study are indirect dependencies (associations) caused by environmental factors. For most microbial association networks, an edge indicates one of the following three alternatives:

1. ecological interaction between two microorganisms,
2. similar or contrary dependence (i.e., preference) to environmental factor/s or a third microorganisms,
3. association by chance.

Indirect associations occur when two microorganisms are both dependent on an abiotic environmental factor (e.g., same nutrients and temperature requirements) or biotic factor (e.g., same prey or predator), but do not interact with one another. Here, indirect association describes the computational effect of indirect dependencies, and observing an association when in fact there is none.

Removing indirect dependencies including environmental effects

To distinguish between direct and indirect interactions, several network construction tools use a probabilistic graphical model (Kurtz *et al.*, 2015; Yang *et al.*, 2017), e.g. SPIEC-EASI (Kurtz *et al.*, 2015, 2019), miic (Verny *et al.*, 2017), or FlashWeave (Tackmann *et al.*, 2019). FlashWeave can also integrate metadata to avoid indirect associations driven by environmental factors but currently does not support missing data. The tool ARACNE (Margolin *et al.*, 2006) aims to eliminate indirect associations by using an information theoretic property (the *Data Processing Inequality*, DPI, in Methods). The extension TimeDelay-ARACNE (Zoppoli *et al.*, 2010) tries to extract dependencies between different times. Another approach including time-delay is implemented in the tool MIDER (Villaverde *et al.*, 2014), which combines mutual information-based distances and entropy reduction to detect indirect interactions (*Mutual Information*, MI, in Methods). PREMER (Villaverde *et al.*, 2018), a successor of MIDER, allows to include previous knowledge, e.g., known non-existent associations.

There are also several prior network construction approaches to reduce indirect associations, e.g., a high prevalence filter that preserves microorganisms present in many samples (Pascual-García *et al.*, 2014). However, this will keep generalist while removing specialist. Another approach divides datasets displaying a great environmental heterogeneity into sub datasets of similar environmental conditions (Röttjers & Faust, 2018). For example, a previous work (Mandakovic *et al.*, 2018) constructed two networks representing bacterial soil communities from two different sections of a pH, temperature, and humidity gradient. Another work (Lima-Mendez *et al.*, 2015) constructed ocean depth-specific networks to account for environmental differences between the surface layer and the deep chlorophyll maximum layer. In addition to dividing samples, an algorithm aiming to correct for habitat filtering effects (Brisson *et al.*, 2019), subtracts, for a given habitat, the mean abundance from each microorganisms within each sample. However, this approach is limited to the identified habitat groups that should have a similar sample size.

In contrast, there are methods accounting for indirect dependencies after network construction. For instance, global silencing, (Barzel & Barabási, 2013) and network deconvolution (Feizi *et al.*, 2013) aim to recover true direct associations from observed correlations. Both techniques are sensitive to missing variables (Alipanahi & Frey, 2013). Another method, called *Sign Pattern*, SP, uses environmental triplets (Lima-Mendez *et al.*, 2015). An environmental triplet contains two microorganisms and one environmental factor, which are associated to each other. SP combines the signs of association scores (positive or negative) to determine if a microbial association should be classified as indirect (SP in Methods). Its major drawback is edge removal where microorganisms with similar environmental preference interact. Along SP and network deconvolution, the *Interaction Information*, II, was applied in (Lima-Mendez *et al.*, 2015). Within an environmental triplet, the II method aims to indicate whether an edge is due entirely to shared environmental preferences (II<0) or whether environmental preferences and true interactions are entangled (II>0). However, II cannot determine which associations in a triplet is indirect (II in Methods). Here, we study several indirect edge detection methods: SP, *Overlap*, (OL, developed here), II, DPI, and their combination.

EnDED is an implementation of four methods and their combination

This article presents EnDED, which implements four approaches, and their combination, to indicate environmentally-driven (indirect) associations in microbial networks. The four methods are: Sign Pattern (Lima-Mendez *et al.*, 2015), Overlap (developed here), Interaction Information (Lima-Mendez *et al.*, 2015; Ghassami & Kiyavash, 2017), and Data Processing Inequality (Cover & Thomas, 2001; Margolin *et al.*, 2006). SP requires an association score that represents co-occurrence when it is positive, and mutual-exclusion when it is negative.

OL requires temporal data with a known start and end of the association to determine whether the microbial association occurs in a time window when both microorganisms are associated to the same environmental factor. The II method indicates the existence of one indirect dependency between three components that are associated with each other. The DPI method states that the association with the smallest mutual information is the indirect association. Here, we evaluate each method and their combination on how well they detect environmentally-driven associations on association networks from simulated data including two environmental factors. Combining methods in an intersection approach retains more true interactions than each method on its own. A union approach was discarded because it would have retained the smallest number of true interactions. We are able to disentangle and filter environmentally-driven edges from microbial association networks (0.95-0.96 in positive predictive value and 0.35-0.83 in accuracy). We also applied EnDED to disentangle and filter environmentally-driven edges from a real marine microbial association network based on ten years of monthly sampling including ten environmental factors. EnDED contributed to both, generating more reliable hypotheses on microbial interactions, and facilitating network analysis by removing edges from dense “hairball” networks. EnDED is publicly available (Deutschmann, 2019).

Results

Simulated data

To evaluate EnDED’s performance in removing environmentally-driven associations, we simulated 1000 abundance time-series datasets with 50 microorganisms and known true interactions between them. We obtained another 1000 datasets with noise (hereafter dwn). We constructed the networks (hereafter simulated networks) with the tool eLSA (Xia *et al.*, 2011, 2013) (see methods). The simulated networks contained on average (computed as the median) 50 nodes and 1087 edges (1063 dwn), of which 60 (59 dwn) were true interactions (edges present in the inferred and true network) and 1026 (1005 dwn) false associations (edges present in the inferred but absent in the true network). Networks inferred from simulated data without noise contained on average one more true interaction but also 21 more false interactions than the networks inferred from simulated data with noise.

A simple approach to discriminate true interactions (desired) from false associations (undesired) would be to use a threshold for the association strength, which could be suitable if the values for true interactions and false associations are i) following different distributions, and ii) the distributions are mainly non-overlapping. We tested the former requirement with a two-sample Kolmogorov-Smirnov test with the R (R Core Team, 2019) function *ks.test*. Using a 95% (99%, 99.9%) confidence level, the distributions were significantly different for 358 (192, 66) simulated datasets and 355 (173, 68) simulated datasets with noise, which is slightly more than one third of them. This indicates that an association strength cut-off is unsuitable to separate true interactions from false associations. More sophisticated approaches than a simple threshold include the methods implemented in EnDED: SP, OL, II, DPI, and their combination.

Combining the methods in an intersection approach (hereafter referred to as intersection combination), we classified on average 348 (228 dwn), that is 32% (22% dwn) of the associations, to be environmentally-driven. The number of correctly detected false associations was on average 332 (219 dwn), i.e., 96% of the removed edges. The resulting networks contained on average 737 (828 dwn) edges. When each method was individually applied more edges were removed: 87% (86% dwn) for SP and OL, 67% (60% dwn) for II, and 44% (32% dwn) for DPI. The fraction of correctly removed edges for individual methods was on average 95%. Comparing the methods on correctly detected false associations, the greatest agreement was observed between SP and OL, whereas DPI appeared to be the most conservative in not agreeing with other methods and, subsequently, reducing the number of detected edges in the intersection combination approach (Supplementary Table S1).Individual methods removed more edges from the network than the intersection combination, where all methods must agree. However, a method's performance is not solely determined by the number of removed edges.

To evaluate the removal of environmentally-driven edges, we scored the different approaches based on five evaluation measurements (see Methods): the true positive rate, TPR, true negative rate, TNR, false positive rate, FPR, positive predicted value, PPV, and accuracy, ACC, (Figure 1 and Supplementary Table S2). In order to determine these measurements, we first determined true and false positives, as well as true and false negatives. A true positive is a false association in the network that is correctly removed by a method, and a false negative is a false association that is incorrectly not removed. A false positive is a true interaction in the network that is incorrectly removed by a method, and a true negative is a true interaction that correctly is not removed by a method. The ideal method maximizes true positives and true negatives and minimizes false positives and false negatives.

The intersection combination under-performed compared to each individual method, SP and OL perform best, and II performs better than DPI according to TPR, FPR and ACC (Figure 1). However, applying each method individually has the drawback of removing more true interactions. On average there are 60 (59 dwn) true interactions in the simulated networks. The individual methods removed 86% (85% dwn) (SP), 85% (84% dwn) (OL), 60% (51% dwn) (II), and 38% (28% dwn) (DPI). Therefore, although the intersection combination removed fewer edges, it outperformed the others according to the TNR because it eliminated fewer of the true interactions, 25% (16% dwn). All methods had high PPV values with half of all measured PPV above ≈0.95. According to PPV, intersection combination performed best and SP and OL performed worst (Figure 1).

Real data

After testing EnDED's performance on simulated networks, we applied it to a real microbial association network, which was constructed from 10 years of monthly samples from January 2004 to December 2013 at the Blanes Bay Microbial Observatory (BBMO) (Gasol *et al.*, 2016). These samples included bacteria and eukaryotes of two size-fractions: picoplankton (0.2-3 µm) and nanoplankton (3-20 µm). We estimated community composition via metabarcoding of the 16S and 18S rRNA gene, and inferred an association network, hereafter referred to as BBMO network (see Methods). The BBMO network contained 762 nodes including 754 ASVs and eight of the ten available environmental factors, and 30498 edges including 29820 microbial edges and 607 edges between a microorganism and an environmental factor. The network contained more positive (24458, 82.0%) than negative (5362, 18.0%) microbial associations (Figure 2).

We found that 25230 (84.6%) of the network edges were in at least one and in maximum six environmental triplets (Figure 2 and Supplementary Table S3). Overall, we detected 35166 environmental triplets within the BBMO network. Of the ten considered environmental factors, PO_4^{3-} and salinity were not associated to any microorganism in the network, and turbidity and NH_4^+ were not found within a triplet. Thus, six environmental factors remained: Temperature (1831 environmentally-driven edges were removed due to Temperature) and day length (652 removed edges) were the top two environmental factors affecting microbial associations, followed by total chlorophyll (175), SiO_2 (5) and NO_3^- (1); no edge was removed due to NO_2^-.

The intersection combination removed 2488 (≈8.3%) associations from the BBMO network. We classified and quantified these indirect edges according to the domain of the nodes (bacteria - eukaryotes, nanoplankton – picoplankton), environmental factor, and the number of triplets a microbial edge was in (Figure 2 and Supplementary Table S4). Compared to the intersection combination, each method individually removed more edges: 84.6% (SP and OL removing all microbial edges present in a triplet), 25.7% (II), and 24.8% (DPI); that is, removal was 3 to 10 times larger.

We also determined for each association the Jaccard index, which indicates how often two microorganisms appear together in the dataset. We assumed that two microbes that appear together < 50% of the time are less likely to have true contemporary ecological interactions and the corresponding association is more likely to be false. We found that only 27.7% of the indirect associations had a Jaccard index above 0.5 compared to 61.1% of the associations that were not indirect. This discrepancy was bigger for negative edges, with 1.2% above and 98.8% below 0.5 (Table 1). The fact that over 72.3% of environmentally-driven associations had a Jaccard index equal or below 0.5 strengthened the decision of their removal.

The intersection combination removed more negative than positive edges, 1554 and 934, respectively (Figure 2). However, there were 20334 positive and 4896 negative microbial associations that were found in at least one environmental triplet, so the method removed 31.7% of the negative and only 4.6% of the positive edges. If we randomly removed 2488 edges, we would expect 18.0 % to be negative (i.e. 448) and 82.0 % of them to be positive (i.e. 2040). If we restrict these calculations to the 25230 microbial associations that were found in at least one environmental triplet, with 20334 of them being positive and 4896 being negative, we would expect to remove 19.4% (i.e. 483) of negative and 80.6% (i.e. 2005) of positive edges. The probability of randomly removing less positive than negative associations is nearly zero, since it follows a multivariate hypergeometric distribution:

$$P(k_{neg}, k_{pos}) = \frac{\binom{N_{neg}}{k_{neg}} \cdot \binom{N_{pos}}{k_{pos}}}{\binom{N}{n}}, \quad \text{Eq. (1)}$$

where N_{pos} and N_{neg} are the number of positive and negative associations in the network, respectively, k_{pos} is the number of removed positive and k_{neg} the removed negative associations from the network, N is the number of associations in the network, and n is the number of removed associations from the network. The removal of more negative edges through intersection combination indicates that this removal was not random or, in other words, that negative associations are more likely to represent environmentally-driven edges.

To evaluate the performance of EnDED on the BBMO network, we considered interactions described in literature and collected in the Protist Interaction Database (PIDA) (Bjorbækmo *et al.*, 2019). Studies typically compare the associations of a network to those reported in the literature at the genus level (Lima-Mendez *et al.*, 2015). The ambiguity in taxonomic classification and the large number of edges challenged this comparison. Thus, we implemented a function to compare strings and match the taxonomic classification of a microorganism in the BBMO network to those in the scientific literature (PIDA). We found that only 29 (0.1%) associations were supported by interactions described in the literature (Table 2). That is, 99.9% of associations in the BBMO network (before applying EnDED) could not be used to evaluate EnDED's performance. These 29 associations describe eight unique interactions between eight microorganisms, and 18 edges were in an environmental triplet to which each method as well as their combination were applied (summary in Table 2). Ideally none of these described associations should be removed by EnDED. Yet, the intersection combination removed five associations (Table 2). In contrast and even worse, SP and OL removed all 18 edges, II eight and DPI nine edges. The additionally removed edges by individual methods are associations between a diatom (*Thalassiosira*) and an unknown *Flavobacteriia*. Considering only the genus level, there were 171 unique genera in the BBMO network, and 700 in PIDA, combined there were 837 microbial genera, and 34 genera in both. Thus, 19.9% of the microbial genera found in the BBMO network were also in PIDA, and 4.9% of the genera found in PIDA were also found in the BBMO network.

Discussion

Using EnDED to disentangle environmental effects in microbial association networks

EnDED makes several indirect-edge removal techniques accessible to microbial ecologists without requiring previous programming experience. These techniques can be used individually or combined. In addition, this work systematically evaluated the different techniques and their combination to remove indirect edges from microbial association networks. Here, we tested only the union and intersection combination of all four methods, but other combination strategies are possible with EnDED. EnDED requires data of the environmental factors in order to predict if an association is environmentally-driven. This is a limitation, since it may be impossible to consider all environmental factors (Lv *et al.*, 2019). However, EnDED can perform well if the major environmental factors, such as, e.g., temperature and nutrient concentrations for marine microorganisms, are provided. Moreover, knowledge of microbial interactions in nature is rather limited and therefore determining the performance of EnDED for real networks is challenging and carries some degree of uncertainty. Thus, EnDED's results should be interpreted with care.

For the simulated networks, we found that each method individually removed on average a moderate to high number of edges. The intersection combination removed fewer edges but kept more true interactions. To understand the impact of the environment, Röttjers and Faust simulated an increasing environmental influence and observed a decrease in retrieving true interactions from inferred associations (Röttjers & Faust, 2018). The observation holds for several network construction methods for cross-sectional data, including CoNet (Faust *et al.*, 2012), SparCC (Friedman & Alm, 2012), SPIEC-EASI (Kurtz *et al.*, 2015), and Spearman correlations. In agreement with these findings, we observed a slight increase in retrieving true interactions when removing environmentally-driven associations in our simulation networks.

In our BBMO dataset, the intersection combination removed a modest number of the edges—a much higher fraction of negative than positive edges. We argue that several negative associations are probably due to different environmental preference (different niches) of microorganisms. The Jaccard index representing a level of microbial co-occurrence, scored equal or below 50% for most negative associations. These may partially represent microorganisms adapted to different seasons. Previous work on the eukaryotic pico- and nano-plankton at the BBMO, using the same basal 10-year dataset used here, indicated a strong seasonality at the community level (Giner *et al.*, 2019).

Comparisons of indirect edge detection on other datasets

In our BBMO network we found that the majority (84.6%) of the microbial edges was within at least one environmental triplet. This was 2.6 times higher than what was found for an association network inferred from data considering microorganisms and small metazoans from two ocean depths across 68 stations around the world and various size fractions (hereafter global interactome) (Lima-Mendez *et al.*, 2015). This global interactome contains 29912 (32.3%) edges that were within at least one environmental triplet (Lima-Mendez *et al.*, 2015). In the previous study, 29900 edges in the global interactome (≈100% of triplets and 32% of all edges) were attributed to environmental factors by SP, similarly to this study as SP removed all edges within triplets in the BBMO network. II indicated 11043 environmentally-driven edges in the global interactome (≈37% of triplets and 12% of all edges) with p-value below 0.05 in a permutation test with 500 iterations. In comparison, II removed a higher fraction of edges in the BBMO network when considering all edges (25.7%), but less when considering within the triplets (30.4%). Network deconvolution suggested 22439 environmentally-driven edges (≈75% of triplets and 24% of all edges) within the global interactome, and the three methods agreed for 8209 edges (≈27% of triplets and 8.9% of all edges). In comparison, we detected slightly less environmentally-driven associations for the BBMO network (8.3% of all edges). These differences suggest that a higher environmental heterogeneity in the dataset may induce more indirect edges. Also, the effects of indirect dependencies may depend on dataset type (e.g., temporal vs. spatial). These possible differences and their effect on environmentally-driven edges should be further investigated.

Using II for the BBMO network, we identified a moderate number of environmentally-driven associations. DPI also identified a moderate number (24.8%, 29.3% when considering only triplets), whereas SP or OL identified a ubiquitous number of environmentally-driven edges (84.6%, 100% when considering only triplets). This indicates that SP and OL are strict and should be used in combination with other methods in an intersection approach.

In another study, the tool FlashWeave (Tackmann *et al.*, 2019) predicted direct microbial interactions in the human microbiome using the Human Microbiome Project (HMP) dataset, including heterogeneous microbial abundance data of 68818 samples (The Human Microbiome Project Consortium: Huttenhower *et al.*, 2012). The inferred networks (with and without metadata) were sparser than our networks. The network with metadata contained 10.7% fewer associations compared to the network without metadata, slightly more than in our results from BBMO.

Factors causing indirect microbial associations

From the simulated networks, we found that using the intersection combination instead of each method individually, we maintained more true interactions at the cost of more false associations in the network—more when considering simulated networks including noise. Comparing our simulated network against the BBMO network, the intersection combination classified a higher number of edges as environmentally-driven in the simulated networks 32% (22% dwn) than in the BBMO network (8.3%). For the simulated data, we previously knew the environmental factor influencing pairwise microbial associations. For the BBMO data, we used ten available environmental factors, but not all factors that could affect microbial dynamics. Even though the most important factors influencing microbial seasonal dynamics at BBMO were considered (Giner *et al.*, 2019), there are several factors that were not measured and that could generate indirect edges. The indirect edges associated to these factors were not detected in our analyses. Similarly, indirect edges associated to biotic interactions (e.g., two bacteria sharing a positive edge as they are symbionts in the same protists) were not considered. Future sampling for microbial interaction research should expand metadata collection in order to detect (more) abiotic and biotic factors that could generate indirect edges.

While temperature and day length (hours of light) were the top two environmental factors affecting microbial associations in the BBMO network, the most important environmental factors in the global interactome (Lima-Mendez *et al.*, 2015) were phosphate concentration and temperature, followed by nitrite concentration and mixed-layer depth. Although we considered PO_4^{3-} and salinity, they were not associated to any microorganism in the network, which may reflect the low variation of these environmental factors in the studied marine site (BBMO). For instance, the standard deviation in the BBMO dataset was < 1 for PO_4^{3-} and salinity, in contrast to the global interactome dataset (Lima-Mendez *et al.*, 2015), where it was about 20-30 when considering all samples. During the Malaspina-2010 Circumnavigation Expedition, the concentrations of trace metals were determined for 110 surface water samples (Pinedo-González *et al.*, 2015). The previous study indicates relationships between primary productivity and trace nutrients, more specifically for the Indian Ocean Cd, the Atlantic Ocean Co, Fe, Cd, Cu, V and Mo, and the Pacific Ocean Fe, Cd, and V. Thus, trace metals are further environmental factors that may play an important role in regulating oceanic primary productivity.

Limitations of EnDED

EnDED detects and removes environmentally-driven indirect edges. However, its triplet analysis could be extended to remove indirect edges driven by taxa, as done with gene triplets (Margolin *et al.*, 2006). A recent update of the network construction tool eLSA (Xia *et al.*, 2011, 2013) permits to examine how a factor, such as a microorganism or environmental variable, mediates the association of two other factors (Ai *et al.*, 2019), which allows the study of interactions between three factors. Furthermore, triplets limit the study to first-order indirect dependencies, neglecting higher-order indirect dependencies. Such limitation was solved for the DPI method by examining associations in quadruplets, quintuplets, and sextuplets (Jang *et al.*, 2013). Implementing higher-order DPI and adjusting the other three methods to account for higher-order indirect dependencies may be promising but one needs to be aware that incorporating higher-order dependencies will also increase the risk of over-fitting. Further, all relevant (measured) environmental factors could be incorporated into the calculation of II, which would combine environmental triplets. However, we reason that such adjustments would require a larger sample size. Both II and DPI calculate MI that measures the dependence between two random variables. EnDED is limited by including one function to estimate the MI. A comparison of four different MI estimates revealed that obtaining the true value of MI is not straightforward, and minor variations of assumptions yield different estimates (Fernandes & Gloor, 2010). Lastly, the conditional mutual information, CMI, which quantifies nonlinear direct relationships among variables, can be underestimated if variables have tight associations in a network (Zhao *et al.*, 2016). The so-called part mutual information, PMI, measurement can help overcome CMI's underestimations. Although using PMI instead of CMI looks promising, calculating PMI is computationally more demanding (Zhao *et al.*, 2016).

Future Perspectives

In this study, we have shown that EnDED with an intersection combination approach provides less dense networks, but still with many potential interactions. We observed a trade-off comparing single methods with the combination approach (intersection combination). Although the latter kept more true interactions, it kept also more false associations. Inferring emergent properties is a key task in microbial ecology to characterize microbial ecosystems from a network-perspective. Thus, if the study aim is to explore patterns of network topology rather than single edges, inferring a network comparable to the real interaction network may be more useful than accuracy of single edges. However, investigations aiming to provide potential interaction partners may use EnDED with the intersection combination approach (e.g., (Latorre *et al.*, 2021)). Specific associations may be validated with experiments or microscopy (Lima-Mendez *et al.*, 2015; Krabberød *et al.*, 2017). However, we suggest to first further reduce the set of potential interaction hypotheses. To improve the selection of interaction hypotheses, we propose to score associations based on re-occurrence: in time, as done with microbial abundance seasonality (Giner *et al.*, 2019), or space, where an association appears in different networks based on different datasets, or different regions of the world. In a previous study using 313 samples, including seven size-fractions, four domains (Bacteria, Archaea, Eukarya, and viruses), and two depths from 68 stations across eight oceanic provinces, 14% of the 81590 predicted biotic interactions were identified as local (Lima-Mendez *et al.*, 2015). Thus, re-occurrent associations may suggest a higher likelihood that the association represents a true ecological interaction, reducing the number of interaction hypotheses to the strongest ones. Another strategy to shortlist interaction hypotheses is to incorporate additional data into the network and use a multi-layer network approach. Such data could be environmental preferences such as temperature or salinity optima, size of cells, presence of chloroplasts, or data obtained from High-Throughput Cultivation (Faust, 2019), microbial community transcriptomes that reveal metabolic pathways (McCarren *et al.*, 2010), or interactions inferred from Single-Cell genome data

(Yoon *et al.*, 2011; Krabberød *et al.*, 2017).

Conclusion

In this paper, we presented EnDED, an analysis tool to reduce the number of environmentally induced indirect edges in inferred microbial networks. Applying EnDED on simulated networks indicated that false associations, driven by environmental variables instead of true interactions, were ubiquitous. However, EnDED's intersection combination classified a minority of associations as environmentally-driven in a real (BBMO) network. Depending on the single method used, we classified a moderate to high number of associations as environmentally-driven in the same network. Nevertheless, associations driven by environmental factors must be determined and quantified to generate more accurate insights regarding true microbial interactions. EnDED provides a step forward in this direction.

Methods

Simulated dataset: time series based on an adjusted generalized Lotka-Volterra model

To evaluate the performance of EnDED, we simulated a time series using an adjusted version of the standard *generalized Lotka-Volterra model*, gLV (Berry & Widder, 2014; Bashan *et al.*, 2016). The gLV can describe the dynamics of microbial communities, by including a first order approach of the microbial interactions. The model's simplicity arises from the assumption of linear interactions, which facilitates implementation and allows fast numerical simulations. The gLV has, however, several limitations (Gonze *et al.*, 2018). For example, gLV neglects higher-order interactions and the additivity of interaction strengths is a weakness because they may be combined in different ways. Also, interactions are often assumed to be constant parameters, but a reducing level of a nutrient may weaken cross-feeding relationships. Moreover, gLV omits the influence of environmental factors, which, for example, can induce oscillations in natural communities (Benincà *et al.*, 2011). Using a model that accounts for nutrients (Kettle *et al.*, 2018) is more realistic but also more complex. More elaborate mechanistic models of microbial dynamics than gLV solve explicitly the global cycling of nutrients and are coupled to the oceanic circulation (see (Vallina *et al.*, 2019) for a review), but the added complexity can hamper understanding about the ecological interactions among microorganisms when compared to a simpler gLV approach. Thus, we chose to use a simpler extension of the gLV to account for the influence of environmental factors (Stein *et al.*, 2013; Dam *et al.*, 2016). In order to allow the growth rates to vary when the environmental variables change, environmental variables can be incorporated directly into the gLV (Dam *et al.*, 2016; Röttjers & Faust, 2018). We simulated a time series using the Klemm-Eguíluz algorithm (Klemm & Eguíluz, 2002), and an adjusted gLV. We adjusted the model by defining microbial growth rates as a function dependent on one seasonal abiotic environmental factor, and added an abiotic environmental factor in the interaction matrix. We then used the time series generated by the gLV to obtain temporal microbial abundance data. With this simulated data, we inferred a network that contained environmentally-driven associations, needed to evaluate the performance of EnDED. We repeated this procedure 1000 times to obtain a large set of simulated networks, and then used the determined abundance tables and Poisson distribution to obtain another 1000 simulated networks including noise. The addition of noise was done by randomly drawing an abundance from the Poisson distribution with λ equaling the original abundance of a specific microorganisms to a specific time.

Adjusting the gLV

To evaluate EnDED, we simulated a time series of microbial abundances with a gLV including true pairwise interactions between 50 microorganisms and adjusted it by incorporating two environmental factors:

$$\frac{dy(t)}{dt} = y(t)[b + Ay(t)]\,, \qquad \text{Eq. (2)}$$

where t is time, $dy(t)/dt$ is the rate of change of microbial abundances as a column vector, $y(t)$ is the vector of microbial abundance at time t, b is the growth rate vector determined through microorganism's specific growth rate functions that depend on an environmental factor (see equation (4)), and A is the interaction matrix.

Interaction matrix

In the interaction matrix A, each coefficient a_{ji} provides the linear effect that a change in the abundance of microorganism i has on the growth of microorganism j (Novak *et al.*, 2016). We simulated the interaction coefficients a_{ji} with the Klemm-Eguíluz algorithm (Klemm &

Eguíluz, 2002), which generates a modular and scale-free matrix. We also set the interaction probability to 0.01, the percentage of positive coefficients to 30%, and diagonal coefficients to zero. Negative diagonal coefficients a_{ii} (i.e., the interaction of a microorganism with itself) can represent intra-specific competition and provides the carrying capacity for each microorganism, preventing its explosive growth (Haydon, 1994). We set the diagonal coefficients $a_{ii} = -0.5$ to avoid excessive microbial abundances in the simulations.

Two abiotic environmental factors

We adjusted the gLV by including two environmental factors. For simplicity, we assume no feedback between the microorganisms and the environmental factors. That is, the environmental factors affect the growth of the microorganisms but not vice-versa. The first environmental factor affects the specific growth rate of each microorganism by interacting with two of their traits: optimal environmental value for growth and tolerance range of environmental values. We simulated the environmental factor using a periodic sinusoidal function (see equation (3)), rounded to 3 digits:

$$\epsilon(t) \triangleq round(\sin(\omega \cdot t), \text{digits} = 3)\,, \qquad \text{Eq. (3)}$$

where t is the time axis (months), $\omega = (-2\pi/T)$ is the signal frequency (radians) and $T = 12$ is the signal periodicity (months); resulting in a signal phase shift of $T/4$ (months). While the first environmental factor is considered to be "external" to the microbial community, the second environmental factor is considered to be "internal", and therefore it is included in the interaction matrix. The interaction coefficients between the microorganisms and the second environmental factor were generated by splitting the microorganisms into two groups: the second abiotic environmental factor influenced positively one half and negatively the other half of the microorganisms. We obtained the interaction coefficients from two uniform distributions defined to range between [-0.8, -0.2] and [0.2, 0.8] respectively. As the microorganisms did not influence the abiotic factor, the corresponding interaction coefficients were set to zero.

Species growth rate

The external seasonal abiotic environmental variable affects the growth rate, g, of each microorganism. This dependency is given by:

$$g(t) \triangleq g_{max}^2 \exp\left(-\frac{1}{2}\frac{\left(\epsilon_{opt} - \epsilon(t)\right)^2}{\sigma^2}\right), \qquad \text{Eq. (4)}$$

where $E(t)$ is the environmental parameter that affects the microorganisms growth rate $g(t)$ at time t, g_{max} is the microorganism' specific maximum growth rate that determines the amplitude of the growth-rate curve, ϵ_{opt} is the microorganism' specific optimal environmental value that determines the peak of the growth-rate curve, and σ is the microorganism' specific ecological tolerance (niche width) determining the environmental range in which the microorganism grows, which determines the length (niche spread) of the growth-rate curve. We obtained the two constant parameters g_{max}, and σ for each microorganism from a uniform distribution ranging between 0.3 and 1 to assure positive values. The values ϵ_{opt} were drawn from a uniform distribution ranging between the minimal and maximal value of the seasonal environmental factor. We defined the internal abiotic environmental factor, which is included in the interaction matrix, through the same function with $g_{max} = 0.8$, $\epsilon_{opt} = 0.5$, and $\sigma = 0.5$. Since the growth rates depend on the environmental factor, they vary seasonally. Different microorganisms will grow better or worse at different times of the year following their environmental niches. This will lead to an asynchrony of their growth rate responses to the environment that will translate into an asynchrony of their abundances in time.

Initial abundances

To obtain the microbial abundances in time with the adjusted gLV, we simulated the initial microbial abundances with a stick-breaking process such that abundances add up to 1, using the function bstick (Jackson, 1993; Legendre & Legendre, 2012), and the package vegan (Oksanen *et al.*, 2019). We generated uneven initial microbial abundances without introducing zeros and set the initial value for the internal abiotic environmental factor included in the interaction matrix to 0.001.

Species abundances in time

Once we have set the initial conditions, we simulated microbial abundances over time by solving the equations given in the adjusted gLV (see equation (2)). Start time was 0, end time

49.5, and sample resolution 0.5 resulting in 100 samples. We used the solver function lsoda (Soetaert *et al.*, 2010). The simulated abundances in time were used to construct an association network, which is referred to as the simulated network.

Real dataset: Blanes Bay Microbial Observatory (BBMO) time series

Microbial abundances

Surface water (≈ 1m depth) was sampled monthly from January 2004 to December 2013, at the BBMO in the North-Western Mediterranean Sea (41°40′N 2°48′E) (Gasol *et al.*, 2016). About 6L of seawater were filtered and separated into picoplankton (0.2-3 μm) and nanoplankton (3-20 μm), as described in (Giner *et al.*, 2019). The DNA was extracted using a phenol-chloroform standard method (Schauer *et al.*, 2003), which has been modified by using Amicon units (Millipore) for purification.

Next, community DNA was extracted, and the 18S ribosomal RNA-gene (V4 region) was amplified in (Giner *et al.*, 2019) using the primer pair TAReukFWD1 and TAReukREV3 (Stoeck *et al.*, 2010). The 16S ribosomal RNA-gene (V4 region) was also amplified from the same DNA extracts using the primers Bakt 341F (Herlemann *et al.*, 2011) and 806R (Apprill *et al.*, 2015). Amplicons were sequenced in a MiSeq platform (2x250bp) at the sequencing service RTL Genomics in Lubbock, Texas. Read quality control, trimming, and inference of Operational Taxonomic Units (OTUs) as Amplicon Sequence Variants (ASV) was made with DADA2 v1.10.1 (Callahan *et al.*, 2016) with the maximum number of expected errors (MaxEE), set to 2 and 4 for the forward and reverse reads, respectively.

ASV sequence abundance tables were obtained for both microbial eukaryotes and prokaryotes. We subsampled both tables to the lowest sequencing depth of 4907 reads, with the rrarefy function from the Vegan package in R (Oksanen *et al.*, 2019) , v2.4-2. We excluded 29 nanoplankton samples (March 2004, February 2005, and May 2010 to July 2012) featuring suboptimal amplicon sequencing. In these, we estimated microbial abundances using seasonally aware missing value imputation by weighted moving average for time series as implemented in the R package imputeTS (Moritz & Gatscha, 2017), v2.8.

Dislodging cells or particles and filter clogging can bias the collection of DNA in either small or large organismal size fractions. To reduce the bias, we divided the sequence abundance sum of the nanoplankton by the picoplankton for each ASV appearing in both size fractions and set the picoplankton abundances to zero if the ratio exceeded 2. Likewise, we set the nanoplankton abundances to zero if the ratio was below 0.5.

Taxonomic classification

The taxonomic classification of each ASV was inferred with the naïve Bayesian classifier method (Wang *et al.*, 2007) together with the SILVA version 132 (Quast *et al.*, 2012) database as implemented in DADA2 (Callahan *et al.*, 2016). In addition, eukaryotic microorganisms were BLASTed (Altschul *et al.*, 1990) against the Protist Ribosomal Reference database [PR2, version 4.10.0; (Guillou *et al.*, 2012)]. If the taxonomic assignment for eukaryotes disagreed between SILVA and PR2, we used the PR2 classification. We removed microorganisms identified as either Metazoa, or Streptophyta, plastids and mitochondria. In addition, we removed Archaeas since the 341F primer is not optimal for recovering this domain (McNichol *et al.*, 2021). The resulting microbial sequence abundance table contained microbial eukaryotic and bacterial ASVs. Rare ASVs were removed, i.e., we kept only ASVs present in more than 15% of the samples and with a sequence abundance sum above 100.

Environmental factors

We measured environmental factors that may affect the ecosystem's dynamics. We considered a total of ten contextual abiotic and biotic variables: day length (hours of light), temperature (C◦), turbidity (Secchi depth m), salinity, total cholorophyll (µg/l), ^{4}and inorganic nutrients— PO$_4^{3-}$ (µM), NH$_4^+$ (µM), NO$_2^-$ (µM), NO$_3^-$ (µM), and SiO$_2$ (µM) (Giner *et al.*, 2019). Water temperature and salinity were sampled in situ with a CTD (Conductivity, Temperature, and Depth) measuring device. Inorganic nutrients were measured with an Alliance Evolution II autoanalyzer (Grasshoff *et al.*, 2009). See (Gasol *et al.*, 2016) for specific details on how other variables were measured.

Network construction

We constructed association networks from the simulated and the real microbial abundance tables and environmental parameters using eLSA (Xia *et al.*, 2011, 2013). We included default normalization and a z-score transformation using median and median absolute deviation. We estimated the p-value with a mixed approach that performs a random permutation test if the theoretical p-values for the comparison are below 0.05; the number of iterations was 2000. Although we are aware of time-delayed interactions and that eLSA (Xia *et al.*, 2011, 2013) could account for them, we considered our sampling interval as too large (1 month) for inferring time-delayed associations with a solid ecological basis. Thus, in our study, we focused on contemporary interactions between co-occurring microbes. For the BBMO dataset, the Bonferroni false discovery rate, q, was calculated for all edges from the p-values using the R function *p.adjust* (R Core Team, 2019). Lastly, we used a significance threshold for the p and q value of 0.001 as suggested in other works (Weiss *et al.*, 2016).

Intersection combination of EnDED—Environmentally-Driven Edge Detection methods

EnDED includes four methods: SP, OL, II, DPI (described below) and their intersection combination (an ensemble approach of the four methods). We applied these methods to find environmentally-driven associations of microorganisms that were within an environmental triplet, as in (Lima-Mendez *et al.*, 2015). An environmental triplet is a special case of a closed triplet where one of the nodes corresponds to an environmental factor and the other two nodes correspond to microorganisms. We define the closed triplet, where there is an edge between each pair of three nodes, as $T = \{v, w, f\}$ where v and w are two microorganisms, and f is an environmental component (see Figure 3).

For the intersection combination, all four individual methods must converge to the same solution, i.e., if all methods classify the microbial edge as environmentally-driven, the edge is removed from the network. If a microbial association is within several environmental triplets, at least one of them must indicate the association as environmentally-driven. In sum, the intersection combination retains an association in the network if no triplet classifies the association as environmentally-driven.

Sign Pattern

The SP method (Lima-Mendez *et al.*, 2015) filters environmentally-driven edges from a network in which a positive association score indicates co-occurrence, and a negative association score indicates mutual exclusion. Let s_{vw} be the sign of the association score of the association between v and w (i.e., $s_{vw} = +$ or $s_{vw} = -$). A closed triplet T has eight SP combinations that group into two sets (see Figure 3). If the product of the three association scores is positive, then the SP suggests that the edge between the two microorganisms is environmentally-driven. Otherwise, if the product of the three association scores is negative, SP does not suggest that the association is environmentally-driven.

Overlap

We have developed the OL method to support the SP for temporal data: a microbial edge should be disregarded as environmentally-driven when the associations are misaligned in time. Thus, OL requires the time when the association begins as well as how long the associations lasts, i.e., duration or length of association in time, both determined by the network construction tool eLSA (Xia *et al.*, 2011, 2013). Given an association between v and w, let b_{vw}^{v} be the beginning of the association for v, b_{vw}^{w} the beginning of the association for w, and d_{vw} be the duration of the association between v and w. Although not used in the BBMO network, OL can consider time-delays by assuming that the beginning of the association is the minimum of the two beginnings, $b_{vw} = \min\,(b_{vw}^{v}, b_{vw}^{w})$, and the end of the association is the maximum, $e_{vw} = \max\,(b_{vw}^{v} + d_{vw}, b_{vw}^{w} + d_{vw})$. We indicate two microorganisms with v and w, and the factor by f . The OL method calculates the overlap O of the microbial association with the two microorganism-environment associations through equation (5). As depicted in Figure 3, if O>60%, the microbial association is considered environmentally-driven.

$$O = 100\frac{min\left(e_{vw}, e_{vf}, e_{wf}\right) - \max\,(b_{vw}, b_{vf}, b_{wf})}{e_{vw} - b_{vw}} \qquad \text{Eq. (5)}$$

Mutual Information and Conditional Mutual Information

The method II employs two measurements: MI and CMI. The former is also used by DPI. Thus, before describing the methods, we first describe the two measurements. MI is a measure of the degree of statistical dependency between two variables (Margolin *et al.*, 2006). We first consider $\boldsymbol{v} = v_1, \ldots, v_n$, $\boldsymbol{w} = w_1, \ldots, w_n$, and $\boldsymbol{f} = f_1, \ldots, f_n$ as discrete random variables. The marginal probability of each discrete state (value) of the variable is denoted by $p(v_i) = P\,(\boldsymbol{v} = v_i)$, the joint probability by $p(v_i, w_j)$, and $p(v_i, w_j, f_k)$, and the conditional probability by $p(v_i|f_k)$, and $p(v_i, w_j|f_k)$. To obtain MI, we calculate the entropy of $\boldsymbol{v}$ as

$$S(\boldsymbol{v}) = -\sum_{i=1}^{n} p(v_i)\log\big(p(v_i)\big), \qquad \text{Eq. (6)}$$

and the joint entropy of $\boldsymbol{v}$ and $\boldsymbol{w}$ as

$$S(\boldsymbol{v},\boldsymbol{w}) = -\sum_{i=1,j=1}^{n} p\big(v_i,w_j\big)\log\Big(p\big(v_i,w_j\big)\Big), \qquad \text{Eq. (7)}$$

using the natural logarithm. The MI of $\boldsymbol{v}$ and $\boldsymbol{w}$ is defined through the sum of their entropies subtracted by their joint entropy:

$$\mathrm{MI}(\boldsymbol{v};\boldsymbol{w}) = S(\boldsymbol{v}) + S(\boldsymbol{w}) - S(\boldsymbol{v},\boldsymbol{w}) \qquad \text{Eq. (8)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} p\big(v_i,w_j\big)\log\left(\frac{p(v_i,w_i)}{p(v_i)p(w_j)}\right), \qquad \text{Eq. (9)}$$

with marginal probabilities $p(v_i) = \sum_{j=1}^{n} p(v_i,w_j)$, and $p(w_j) = \sum_{i=1}^{n} p(v_i,w_j)$.

The measurement CMI is the expected value of the MI of two random variables given a third random variable. It is defined as

$$\mathrm{CMI}(\boldsymbol{v};\boldsymbol{w}|\boldsymbol{f}) = S(\boldsymbol{v},\boldsymbol{f}) + S(\boldsymbol{w},\boldsymbol{f}) - S(\boldsymbol{v},\boldsymbol{w},\boldsymbol{f}) - S(\boldsymbol{f}) \qquad \text{Eq. (10)}$$

$$= \sum_{k=1}^{n} p(f_k)\sum_{i=1}^{n}\sum_{j=1}^{n} p\big(v_i,w_j|f_k\big)\log\left(\frac{p(v_i,w_i|f_k)}{p(v_i|f_k)p(w_j|f_k)}\right) \qquad \text{Eq. (11)}$$

$$= \sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} p\big(v_i,w_j,f_k\big)\log\left(\frac{p(f_k)p(v_i,w_i,f_k)}{p(v_i,f_k)p(w_j,f_k)}\right).$$

Interaction Information

The II is calculated with microbial abundance and environmental data. In this study, as in (Lima-Mendez *et al.*, 2015), II is computed as the difference of the CMI and MI:

$$\mathrm{II} = \mathrm{CMI} - \mathrm{MI}. \qquad \text{Eq. (12)}$$

In other works (Ghassami & Kiyavash, 2017), the II is defined with a different sign convention: II = MI − CMI. In our study, if II is positive, the method suggests that the microbial association is not environmentally-driven. If II is negative, there is an environmentally-driven association within the closed triplet. However, this method cannot detect which of the three associations is indirect. In other works (Lima-Mendez *et al.*, 2015), the microbial association is assumed to be environmentally-driven if II is negative, but here we suggest to combine it with DPI (see below).

Significance of Interaction Information

We determined the significance of II following a strategy from (North *et al.*, 2002; Veech, 2012). We used a parameter-free permutation test and computed the p-value by randomizing the environmental vector $\boldsymbol{f}$. Since the MI is independent of the environmental factor and therefore remains constant, the significance of the II is the same as the CMI. Thus, we determined the significance of CMI with 1000 permutations: we randomized the environmental vector $\boldsymbol{f}$ and recalculated the CMI 1000 times, obtaining a CMI_i with $i \in \{1,...,1000\}$. Afterwards, we quantified with c how many random CMI_i were at least as small as the original CMI_i: $c = |i: \mathrm{CMI}_i \leq \mathrm{CMI}_{original}, i \in \{1,...,1000\}|$. We calculated the p-value as

$$p = \frac{c+1}{1000+1}\ . \qquad \text{Eq. (13)}$$

Data Processing Inequality

As mentioned above, the II method can detect if an indirect association exists within a triplet but cannot determine which of the three associations is indirect. Thus, we added DPI to EnDED. DPI states that if two components v and w interact only through a third component f (i.e., in a network v and w are connected through a path containing f and there is no alternative path between v and w), then the MI of v and w, $\mathrm{MI}(\boldsymbol{v};\ \boldsymbol{w})$ is smaller than $\mathrm{MI}(\boldsymbol{v};\ \boldsymbol{f})$ and $\mathrm{MI}(\boldsymbol{w};\ \boldsymbol{f})$ (Cover & Thomas, 2001):

$$\mathrm{MI}(\boldsymbol{v};\boldsymbol{w}) \leq \min\{\mathrm{MI}(\boldsymbol{v};\boldsymbol{f}), \mathrm{MI}(\boldsymbol{w};\boldsymbol{f})\}\ . \qquad \text{Eq. (14)}$$

While DPI has been used in previous works on gene triplets (Margolin *et al.*, 2006), we used the DPI method for environmental triplets. We compared the MI between the two microorganisms with the MI between a microorganism and the environmental factor. If the MI between the microorganisms is the smallest, then the method suggests that the edge is environmentally-driven. This method complements the II method.

Equal Width Discretization

To compute the MI, CMI, and subsequently II, we discretized the abundance data and environmental parameters. EnDED uses the equal width discretization algorithm, which creates equal sized ranges (also called bins or buckets) for an abundance vector $\boldsymbol{v} = (v_1, \dots, v_n)$ between the lowest value (v_{min}) and highest value (v_{max}). It is a procedure implemented in other works (Meyer *et al.*, 2008). Given vector $\boldsymbol{v}$ of length n (that is the sample size) and number of bins $|B| = \lfloor \sqrt{n} \rfloor$, the discretized value v_d of variable v in vector $\boldsymbol{v}$ is:

$$v_d = \left\lceil \frac{(v - v_{min}) \cdot |B|}{v_{max}} \right\rceil . \qquad \text{Eq. (15)}$$

This equation assumes positive values. However, if $\boldsymbol{v}$ contains negative values, $v_{min} < 0$, we adjust equation (15) by substituting v_{max} for $v'_{max} = v_{max} - v_{min}$. This method does not fill in missing values, and it is limited by the presence of outliers as most values would go within the same bin. We can solve this problem with a different discretization method (where bins have the same number of elements) but we have not implemented it in the current version of EnDED.

Applying EnDED to networks constructed from simulated and real data

We applied EnDED to association networks constructed from time series of simulated abundances and estimated microbial abundances from sequence data. The simulated networks were based on a gLV, while the real network was based on data from the BBMO. For the methods II and DPI we also included the corresponding abundance tables, and environmental factors. EnDED was run with the OL threshold of 60%. We set the significance threshold for the II score to 0.05 and used 1000 iterations.

Evaluation of EnDED's performance

Simulated network

We evaluated EnDED with the simulated interaction matrices, which revealed the number of true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) before and after removing associations that were classified as environmentally-driven. We assumed that associations not present in the interaction matrices, are environmentally-driven. We consider P as the number of all false associations, both true positive and false negative detected environmentally-driven edges: $\mathrm{P} = \mathrm{TP} + \mathrm{FN}$, and N as the number of all true interactions, i.e., all true negative and false positive detected environmentally-driven edges: $\mathrm{N} = \mathrm{TN} + \mathrm{FP}$. Then, we calculated the true positive rate (sensitivity), by dividing the number of true positives by the number of all real positives: $\mathrm{TPR} = (\mathrm{TP})/(\mathrm{P})$. Equivalently, we can also calculate the true negative rate (specificity) by dividing the number of true negatives by the number of all real negatives, $\mathrm{TNR} = (\mathrm{TN})/(\mathrm{N})$. The false positive rate (fall out) is the complementary to TNR, i.e. $\mathrm{FPR} = 1 - \mathrm{TNR}$. The positive predictive value (precision) can be calculated by dividing the number of true positives by the sum of all predicted positives, $\mathrm{PPV} = (\mathrm{TP})/(\mathrm{TP} + \mathrm{FP})$. The accuracy is calculated by dividing the sum of true positives and true negatives by the sum of all real positives and real negatives, $\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{P} + \mathrm{N})$.

Real Dataset

Literature based database

The real network evaluation is limited since the true interactions and the microorganisms that do not interact with each other are poorly known. We assessed true interactions known in the literature based on the genus, which are compiled within the Protist Interaction Database, PIDA (Bjorbækmo *et al.*, 2019). On October 15th 2019, PIDA contained 2448 interactions. Although our dataset contains protists as well as bacteria, we were unable to evaluate interactions between bacteria through PIDA.

Jaccard index

In ecology, the Jaccard index (Jaccard similarity coefficient) is normally used for communities. Here, for each pair of microorganisms in the BBMO network, we computed the Jaccard index as the number of samples in which both microorganisms occur, divided by the number of samples in which at least one of the two microorganisms are present.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

EnDED is publicly available: https://github.com/InaMariaDeutschmann/EnDED. This repository contains the file "FromDataSimulationToEvaluatingEnDED.RMD", which contains R code to generate simulated abundance tables, commands to run eLSA network construction and EnDED, as well as the command to run a C++ program (included as well) and R code used for evaluation. The repository folder BBMO data contains the BBMO abundance table, the taxonomic classification table, and the BBMO network including results of EnDED.

Competing interests

The authors declare that they have no competing interests.

Funding

This project and IMD received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 675752 (SINGEK: http://www.singek.eu). RL was supported by a Ramón y Cajal fellowship (RYC-2013-12554, MINECO, Spain). This work was also supported by the projects INTERACTOMICS (CTM2015-69936-P, MINECO, Spain), MINIME (PID2019-105775RB-I00, AEI, Spain) and MicroEcoSystems (240904, RCN, Norway) to RL.

Author's contributions

IMD, GLM, JR, KF and RL designed and conceived the project. IMD performed data analysis, data simulation, and implementation of EnDED. IMD received substantial feedback on established indirect detection methods from GLM and KF, on data simulation from SMV and KF, on network construction from AKK, and on evaluation of EnDED from GLM and KF (measurements for simulation dataset) and AKK (literature based database for real dataset). RL processed the amplicon data from BBMO generating ASV tables. AKK ran the eLSA network construction tool for the BBMO dataset and IMD ran the tool for the simulation datasets. RL provided funding for the project. The original draft was written by IMD. IMD, GLM, AKK, SMV, KF and RL contributed substantially to manuscript revisions. All authors approved the final version of the manuscript.

Acknowledgements

We thank all members of the Blanes Bay Microbial Observatory sampling team and the multiple projects funding this collaborative effort over the years. We also thank collaborators at www.thepapermill.eu for help with critical reading in the early stages of the manuscript. Part of the analyses have been performed at the Marbits bioinformatics core at ICM-CSIC (https://marbits.icm.csic.es).

References

AI, D., LI, X., PAN, H., CHEN, J., CRAM, J.A., & XIA, L.C. (2019) Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis. *BMC Genomics*, **20**, 185.

AITCHISON, J. (1981) A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*.

ALIPANAHI, B. & FREY, B.J. (2013) Network cleanup. *Nature Biotechnology*, **31**, 714–715.

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W., & LIPMAN, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

APPRILL, A., MCNALLY, S., PARSONS, R., & WEBER, L. (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology*, **75**, 129–137.

BARZEL, B. & BARABÁSI, A.-L. (2013) Network link prediction by global silencing of indirect correlations. *Nature Biotechnology*, **31**, 720–725.

BASHAN, A., GIBSON, T.E., FRIEDMAN, J., CAREY, V.J., WEISS, S.T., HOHMANN, E.L., & LIU, Y.-Y. (2016) Universality of human microbial dynamics. *Nature*, **534**, 259–262.

BENINCÀ, E., DAKOS, V., VAN NES, E.H., HUISMAN, J., & SCHEFFER, M. (2011) Resonance of Plankton Communities with Temperature Fluctuations. *The American Naturalist*, **178**, E85–E95.

BERRY, D. & WIDDER, S. (2014) Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*, **5**, 219.

BJORBÆKMO, M.F.M., EVENSTAD, A., RØSÆG, L.L., KRABBERØD, A.K., & LOGARES, R. (2019) The planktonic protist interactome: where do we stand after a century of research? *The ISME Journal*, DOI: 10.1038/s41396-019-0542-5.

BRISSON, V., SCHMIDT, J., NORTHEN, T.R., VOGEL, J.P., & GAUDIN, A. (2019) A New Method to Correct for Habitat Filtering in Microbial Correlation Networks. *Frontiers in Microbiology*, **10**, 585.

CALLAHAN, B.J., MCMURDIE, P.J., ROSEN, M.J., HAN, A.W., JOHNSON, A.J.A., & HOLMES, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, **13**, 581–583.

COVER, T.M. & THOMAS, J.A. (2001) Inequalities in Information Theory. *Elements of Information Theory*.

DAM, P., FONSECA, L.L., KONSTANTINIDIS, K.T., & VOIT, E.O. (2016) Dynamic models of the complex microbial metapopulation of lake mendota. *npj Systems Biology and Applications*, **2**, 16007.

DELONG, E.F. (2009) The microbial ocean from genomes to biomes. *Nature*.

DEUTSCHMANN, I.M. (2019) *EnDED - - Environmentally-Driven Edge Detection Program*. Zenodo.

FALKOWSKI, P.G., FENCHEL, T., & DELONG, E.F. (2008) The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*.

FAUST, K. (2019) Towards a Better Understanding of Microbial Community Dynamics through High-Throughput Cultivation and Data Integration. *mSystems*, **4**.

FAUST, K. & RAES, J. (2012) Microbial interactions: from networks to models. *Nature*

Reviews Microbiology, **10**, 538–550.

Faust, K. & Raes, J. (2016) CoNet app: inference of biological association networks using Cytoscape [version 2; peer review: 2 approved]. *F1000Research*, **5**.

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012) Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Computational Biology*.

Feizi, S., Marbach, D., Médard, M., & Kellis, M. (2013) Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature Biotechnology*, **31**, 726–733.

Fernandes, A.D. & Gloor, G.B. (2010) Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics*, **26**, 1135–1139.

Friedman, J. & Alm, E.J. (2012) Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology*, **8**, 1–11.

Gasol, J.M., Cardelús, C., G Morán, X.A., Balagué, V., Forn, I., Marrasé, C., Massana, R., Pedrós-Alió, C., Montserrat Sala, M., Simó, R., Vaqué, D., & Estrada, M. (2016) Seasonal patterns in phytoplankton photosynthetic parameters and primary production at a coastal NW Mediterranean site. *Scientia Marina*.

Ghassami, A. & Kiyavash, N. (2017) Interaction information for causal inference: The case of directed triangle. *2017 IEEE International Symposium on Information Theory (ISIT)*. pp. 1326–1330.

Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., Gasol, J.M., Logares, R., & Massana, R. (2019) Quantifying long-term recurrence in planktonic microbial eukaryotes. *Molecular Ecology*, **28**, 923–935.

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., & Egozcue, J.J. (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, **8**, 2224.

Gonze, D., Coyte, K.Z., Lahti, L., & Faust, K. (2018) Microbial communities as dynamical systems. *Current Opinion in Microbiology*, **44**, 41–49.

Grasshoff, K., Kremling, K., & Ehrhardt, M. (2009) *Methods of seawater analysis*. John Wiley & Sons.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W.H.C.F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., & Christen, R. (2012) The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, **41**, D597–D604.

Haydon, D. (1994) Pivotal Assumptions Determining the Relationship between Stability and Complexity: An Analytical Synthesis of the Stability-Complexity Debate. *The American Naturalist*, **144**, 14–29.

Herlemann, D.P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J.J., & Andersson, A.F. (2011) Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal*, **5**, 1571–1579.

Jackson, D.A. (1993) Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology*, **74**, 2204–2214.

Jang, I.S., Margolin, A., & Califano, A. (2013) hARACNe: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests. *Interface Focus*, **3**, 20130011.

Kallmeyer, J., Pockalny, R., Adhikari, R.R., Smith, D.C., & D'Hondt, S. (2012) Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences*, **109**, 16213–16216.

Kettle, H., Holtrop, G., Louis, P., & Flint, H.J. (2018) microPop: Modelling microbial populations and communities in R. *Methods in Ecology and Evolution*, **9**, 399–409.

Klemm, K. & Eguíluz, V.M. (2002) Growing scale-free networks with small-world behavior. *Physical Review E*, **65**, 057102.

Krabberød, A.K., Bjorbækmo, M.F.M., Shalchian-Tabrizi, K., & Logares, R. (2017) Exploring the oceanic microeukaryotic interactome with metaomics approaches. *Aquatic Microbial Ecology*, **79**, 1–12.

Kurtz, Z.D., Bonneau, R., & Müller, C.L. (2019) Disentangling microbial associations from hidden environmental and technical factors via latent graphical models. *bioRxiv*, DOI: 10.1101/2019.12.21.885889.

Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., & Bonneau, R.A. (2015) Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*.

Latorre, F., Deutschmann, I.M., Labarre, A., Obiol, A., Krabberød, A.K., Pelletier, E., Sieracki, M.E., Cruaud, C., Jaillon, O., Massana, R., & Logares, R. (2021) Niche adaptation promoted the evolutionary diversification of tiny ocean predators. *Proc Natl Acad Sci USA*, **118**, e2020955118.

Layeghifard, M., Hwang, D.M., & Guttman, D.S. (2017) Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*.

Legendre, P. & Legendre, L.F. (2012) *Numerical ecology*, vol. 24. Elsevier.

Li, C., Lim, K.M.K., Chng, K.R., & Nagarajan, N. (2016) Predicting microbial interactions through computational approaches. *Methods*.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S., Berline, L., Bontempi, G., Cabello, A.M., Coppola, L., Cornejo-Castillo, F.M., d'Ovidio, F., De Meester, L., Ferrera, I., Garet-Delmas, M.-J., Guidi, L., Lara, E., Pesant, S., Royo-Llonch, M., Salazar, G., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Gorsky, G., Not, F., Ogata, H., Speich, S., Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S.G., Sunagawa, S., Bork, P., Sullivan, M.B., Karsenti, E., Bowler, C., de Vargas, C., & Raes, J. (2015) Determinants of community structure in the global plankton interactome. *Science*, **348**, 1262073.

Locey, K.J. & Lennon, J.T. (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, **113**, 5970–5975.

Lv, X., Zhao, K., Xue, R., Liu, Y., Xu, J., & Ma, B. (2019) Strengthening Insights in Microbial Ecological Networks from Theory to Applications. *mSystems*, **4**, e00124-19.

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., Bihouée, A., Jean, G., Díaz, F.P., Fernández-Gómez, B., Cabrera, P., Gaete, A., Latorre, C., Gutiérrez, R.A., Maass, A., Cambiazo, V., Navarrete, S.A., Eveillard, D., & González, M. (2018) Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Scientific Reports*, **8**, 5875.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., & Califano, A. (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7, S7.

McCarren, J., Becker, J.W., Repeta, D.J., Shi, Y., Young, C.R., Malmstrom, R.R., Chisholm, S.W., & DeLong, E.F. (2010) Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proceedings of the National Academy of Sciences*, **107**, 16420–16427.

McNichol, J., Berube, P.M., Biller, S.J., Fuhrman, J.A., & Gilbert, J.A. (2021) Evaluating and Improving Small Subunit rRNA PCR Primer Coverage for Bacteria, Archaea, and Eukaryotes Using Metagenomes from Global Ocean Surveys. *mSystems*, **6**, e00565-21.

Meyer, P.E., Lafitte, F., & Bontempi, G. (2008) minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics*, **9**, 461.

Moritz, S. & Gatscha, S. (2017) *imputeTS: Time Series Missing Value Imputation*.

North, B.V., Curtis, D., & Sham, P.C. (2002) A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. *The American Journal of Human Genetics*, **71**, 439–441.

Novak, M., Yeakel, J.D., Noble, A.E., Doak, D.F., Emmerson, M., Estes, J.A., Jacob, U., Tinker, M.T., & Wootton, J.T. (2016) Characterizing Species Interactions to Understand Press Perturbations: What Is the Community Matrix? *Annual Review of Ecology, Evolution, and Systematics*, **47**, 409–432.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., & Wagner, H. (2019) *vegan: Community Ecology Package*.

Pascual-García, A., Tamames, J., & Bastolla, U. (2014) Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? *BMC Microbiology*, **14**, 284.

Pinedo-González, P., West, A.J., Tovar-Sánchez, A., Duarte, C.M., Marañón, E., Cermeño, P., González, N., Sobrino, C., Huete-Ortega, M., Fernández, A., López-Sandoval, D.C., Vidal, M., Blasco, D., Estrada, M., & Sañudo-Wilhelmy, S.A. (2015) Surface distribution of dissolved trace metals in the oligotrophic ocean and their influence on phytoplankton biomass and productivity. *Global Biogeochemical Cycles*, **29**, 1763–1781.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F.O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–D596.

R Core Team (2019) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Röttjers, L. & Faust, K. (2018) From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews*, **42**, 761–780.

Schauer, M., Balagué, V., Pedrós-Alió, C., & Massana, R. (2003) Seasonal changes in the taxonomic composition of bacterioplankton in a coastal oligotrophic system. *Aquatic Microbial Ecology*, **31**, 163–174.

Soetaert, K., Petzoldt, T., & Setzer, R.W. (2010) Solving Differential Equations in R: Package deSolve. *Journal of Statistical Software*, **33**, 1–25.

Stein, R.R., Bucci, V., Toussaint, N.C., Buffie, C.G., Rätsch, G., Pamer, E.G., Sander, C., & Xavier, J.B. (2013) Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *PLOS Computational Biology*, **9**, 1–11.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breiner, H.-W., & Richards, T.A. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, **19**, 21–31.

Tackmann, J., Rodrigues, J.F.M., & von Mering, C. (2019) Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data. *Cell Systems*, **9**, 286-296.e8.

The Human Microbiome Project Consortium: Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., Giglio, M.G., Hallsworth-Pepin, K., Lobos, E.A., Madupu, R., Magrini, V., Martin, J.C., Mitreva, M., Muzny, D.M., Sodergren, E.J., Versalovic, J., Wollam, A.M., Worley, K.C., Wortman, J.R., Young, S.K., Zeng, Q., Aagaard, K.M., Abolude, O.O., Allen-Vercoe, E., Alm, E.J., Alvarado, L., Andersen, G.L., Anderson, S., Appelbaum, E., Arachchi, H.M., Armitage, G., Arze, C.A., Ayvaz, T., Baker C.C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M.J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G.A., Buhay, C.J., Busam, D.A., Campbell, J.L., Canon, S.R., Cantarel, B.L., Chain, P.S.G., Chen, I.-M.A., Chen, L., Chhibba, S., Chu, K., Ciulla, D.M., Clemente, J.C., Clifton, S.W., Conlan, S., Crabtree, J., Cutting, M.A., Davidovics, N.J., Davis, C.C., DeSantis, T.Z., Deal, C., Delehaunty, K.D., Dewhirst, F.E., Deych, E., Ding, Y., Dooling, D.J., Dugan, S.P., Michael Dunne, W., Scott Durkin, A., Edgar, R.C., Erlich, R.L., Farmer, C.N., Farrell, R.M., Faust, K., Feldgarden, M., Felix, V.M., Fisher, S., Fodor, A.A., Forney, L.J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D.C., Fronick, C.C., Fulton, L.L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M.Y., Goldberg, J.M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Kinder Haake, S., Haas, B.J., Hamilton, H.A., Harris, E.L., Hepburn, T.A., Herter, B., Hoffmann, D.E., Holder, M.E., Howarth, C., Huang, K.H., Huse, S.M., Izard, J., Jansson, J.K., Jiang, H., Jordan, C., Joshi, V., Katancik, J.A., Keitel, W.A., Kelley, S.T., Kells, C., King, N.B., Knights, D., Kong, H.H., Koren, O., Koren, S., Kota, K.C., Kovar, C.L., Kyrpides, N.C., La Rosa, P.S., Lee, S.L., Lemon, K.P., Lennon, N., Lewis, C.M., Lewis, L., Ley, R.E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C.A., Dwayne Lunsford, R., Madden, T., Mahurkar, A.A., Mannon, P.J., Mardis, E.R., Markowitz, V.M., Mavromatis, K., McCorrison, J.M., McDonald, D., McEwen, J., McGuire, A.L., McInnes, P., Mehta, T., Mihindukulasuriya, K.A., Miller, J.R., Minx, P.J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S.M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K.S., Pop, M., Priest, M.E., Proctor, L.M., Qin, X., Raes, J., Ravel, J., Reid, J.G., Rho, M., Rhodes, R., Riehle, K.P., Rivera, M.C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M.C., Russ, C., Sanka, R.K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J.A., Schloss, P.D., Schmidt, T.M., Scholz, M., Schriml, L., Schubert, A.M., Segata, N., Segre, J.A., Shannon, W.D., Sharp, R.R., Sharpton, T.J., Shenoy, N., Sheth, N.U., Simone, G.A., Singh, I., Smillie, C.S., Sobel, J.D., Sommer, D.D., Spicer, P., Sutton, G.G., Sykes, S.M., Tabbaa, D.G., Thiagarajan, M., Tomlinson, C.M., Torralba, M., Treangen, T.J., Truty, R.M., Vishnivetskaya, T.A., Walker, J., Wang, L., Wang, Z., Ward, D.V., Warren, W., Watson, M.A., Wellington, C., Wetterstrand, K.A., White, J.R., Wilczek-Boney, K., Wu, Y., Wylie, K.M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B.P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J.D., Birren, B.W., Gibbs, R.A., Highlander, S.K., Methé, B.A., Nelson, K.E., Petrosino, J.F., Weinstock, G.M., Wilson, R.K., & White, O. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Vallina, S.M., Martinez-Garcia, R., Smith, S.L., & Bonachela, J.A. (2019) Models in Microbial Ecology. *Encyclopedia of Microbiology (Fourth Edition)*, Fourth Edition ed. (Schmidt, T.M. ed). Oxford: Academic Press, pp. 211–246.

Veech, J.A. (2012) Significance testing in ecological null models. *Theoretical Ecology*, **5**, 611–616.

Verny, L., Sella, N., Affeldt, S., Singh, P.P., & Isambert, H. (2017) Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology*, **13**, 1–25.

Villaverde, A.F., Becker, K., & Banga, J.R. (2018) PREMER: A Tool to Infer Biological Networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **15**, 1193–1202.

Villaverde, A.F., Ross, J., Morán, F., & Banga, J.R. (2014) MIDER: Network Inference with Mutual Information Distance and Entropy Reduction. *PLOS ONE*, **9**, 1–15.

Wang, Q., Garrity, G.M., Tiedje, J.M., & Cole, J.R. (2007) Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F., Zhou, J., & Knight, R. (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, **10**, 1669–1681.

WHITMAN, W.B., COLEMAN, D.C., & WIEBE, W.J. (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, **95**, 6578–6583.

WORDEN, A.Z., FOLLOWS, M.J., GIOVANNONI, S.J., WILKEN, S., ZIMMERMAN, A.E., & KEELING, P.J. (2015) Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*, **347**.

XIA, L.C., AI, D., CRAM, J., FUHRMAN, J.A., & SUN, F. (2013) Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics*, **29**, 230–237.

XIA, L.C., STEELE, J.A., CRAM, J.A., CARDON, Z.G., SIMMONS, S.L., VALLINO, J.J., FUHRMAN, J.A., & SUN, F. (2011) Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Systems Biology*, **5**, S15.

XIAO, Y., ANGULO, M.T., FRIEDMAN, J., WALDOR, M.K., WEISS, S.T., & LIU, Y.-Y. (2017) Mapping the ecological networks of microbial communities. *Nature Communications*, **8**, 2042.

YANG, Y., CHEN, N., & CHEN, T. (2017) Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model. *Cell Systems*, **4**, 129-137.e5.

YOON, H.S., PRICE, D.C., STEPANAUSKAS, R., RAJAH, V.D., SIERACKI, M.E., WILSON, W.H., YANG, E.C., DUFFY, S., & BHATTACHARYA, D. (2011) Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists. *Science*, **332**, 714–717.

ZHAO, J., ZHOU, Y., ZHANG, X., & CHEN, L. (2016) Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences*, **113**, 5130–5135.

ZOPPOLI, P., MORGANELLA, S., & CECCARELLI, M. (2010) TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*, **11**, 154.

Figures

Figure 1: **Evaluation of EnDED: intersection combination and individual methods on simulated networks**. Using 1000 simulated networks, and 1000 simulated networks incorporating noise, we evaluated EnDED's performance. Plot A) displays the evaluation measurements true positive rate (TRP), true negative rate (TNR), accuracy (ACC), and positive predictive value (PPV) for each individual method, i.e., Sign Pattern (SP), Overlap (OL), Interaction Information (II), and Data Processing Inequality (DPI), as well as the intersection combination (Combi). SP and OL perform best according to TRP and ACC, while the intersection combination performs best according to TNR. All methods performed well according to PPV. The intersection combination, DPI and II performed better on noisy data according to TNR because less edges were removed along with less true interactions. Plot B) displays the ROC curve for each environmentally-driven edge detection method as well as their intersection combination.

Figure 2: **Quantification of environmentally-driven associations in the BBMO network.** For A) the first column shows the number and fraction of microbial associations divided by domain: Bacteria-Bacteria associations (B), Bacteria-Eukaryote associations (BE), and Eukaryote-Eukaryote associations (E). The second column shows the number and fraction of associations divided by size-fractions: association within the nano size fraction (n), within the pico size fraction (p), and between these two size fractions (np). The third column shows all microbial edges connected to an environmental parameter: Temperature (Tem), Day length (Day), Chlorophyll (Chl), inorganic nutrients NO_3^- (NO3), SiO_2 (Si), and NO_2^- (NO2). The last column shows the number and fraction of edges divided in how many triplets they have been found ranging from no triplets (0) to six triplets. The first two rows display the number and fraction of microbial associations of the BBMO network before applying EnDED. Positive associations are indicated with black, negative associations with red. The last two rows indicate in blue the fraction of environmentally-driven edges among the positive (third row) and negative (fourth row) microbial associations. B) The left network shows in black the positive and in red the negative associations. The right network shows the number of triplets a microbial edge is in ranging from one (green) to six (orange), and no triplet (black). The middle network shows in blue the environmentally-driven associations that were detected by the intersection combination of the four methods Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality.

Figure 3: **EnDED Methods Overview**. EnDED is an implementation of four methods aiming to determine whether an edge between two microorganisms is indirect through the action of an environmental factor. The four methods are: Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality (see Methods). Each method can be used individually or in combination. Here, we show the intersection combination approach, i.e., only if all methods classify an edge as indirect, it is removed from the network. Otherwise, the edge is classified as not indirect and kept in the network.

Tables

Table 1: **Jaccard index of edges.** The BBMO network before applying EnDED contained 29820 edges of which 2488 (8.3%) were environmentally-driven (indirect). Considering the Jaccard index for these indirect edges, 688 (27.7% of indirect edges) score above 50%, and 1800 (72.3%) score below or equal to 50%. In contrast, 61.1% of edges not considered as indirect have a Jaccard index above 50%, and 38.9% of all not indirect edges have a Jaccard index equal or below 50%.

	All edges	Jaccard index>50	Jaccard index≤50
BBMO network	29 820 (100%)	17 383 (58.3%)	12 437 (41.7%)
positive edges	24 458 (82.0%)	17 212 (70.4%)	7 246 (29.6%)
negative edges	5 362 (18.0%)	171 (3.2%)	5 191 (96.8%)
indirect (intersection)	2 488 (8.3%)	688 (27.7%)	1 800 (72.3%)
positive + indirect (intersection)	934 (3.1%)	670 (71.7%)	264 (28.3%)
negative + indirect (intersection)	1 554 (5.2%)	18 (1.2%)	1 536 (98.8%)
not indirect (all)	27 332 (91.7%)	16 695 (61.1%)	10 637 (38.9%)
not indirect (min 1 triplet)	22 742 (76.3%)	14 242 (62.6%)	8 500 (37.4%)
not indirect (no triplet)	4 590 (15.4%)	2 453 (53.4%)	2 137 (46.6%)
Sign Pattern	25 230 (84.6%)	14 930 (59.2%)	10 300 (40.8%)
Overlap	25 230 (84.6%)	14 930 (59.2%)	10 300 (40.8%)
Interaction Information	7 672 (25.7%)	4 962 (64.7%)	2 710 (35.3%)
Data Processing Inequality	7 394 (24.8%)	1 862 (25.2%)	5 532 (74.8%)

Table 2: **Interactions found in the BBMO network that have been reported in the literature.** The table mentions whether or not the associations were removed or kept by EnDED via the combination interaction approach. For example, the association between the ASVs classified as Dia. Thalassiosira and ASVs classified as F. unknown Flavobacteriia has been found 17 times in the network: 4 were removed and 13 were kept.

Microorganisms	EnDED	ID in PIDA
Included in 1, 2, 3, or 4 triplets		
Dia. Thalassiosira - Dino. Heterocapsa	1 removed	1665
Dia. Thalassiosira - F. unknown Flavobacteriia	4 removed 13 kept	2199
Not included in a triplet		
Dino. Heterocapsa - Dino. Prorocentrum	1 kept	1501, 1511
Dino. Gyrodinium - Dino. Heterocapsa	1 kept	1313, 1314, 1780, 1783
Dino. Prorocentrum - Dino. Gymnodinium	2 kept	1499
Dino. Prorocentrum - Dino. Prorocentrum	4 kept	1509, 1510
Dino. Prorocentrum - Dino. Scrippsiella	2 kept	1513
F. unknown Flavobacteriia - Dia. Pseudo-nitzschia	1 kept	2196

Abbreviations indicate Dia - Diatomea; Dino - Dinoflagellata; C - Ciliophora; F - Flavobacteriia; ID in PIDA refers to the number PIDA gave to an interaction described in the literature.

Supplementary Material

Supplementary Table S1: **Comparison between methods on correctly detecting false associations.** We computed the fraction (in percentage) of correctly detected false associations for each of the 1000 simulated datasets. There are only few edges that are detected by only one approach (first four rows). The most prominent groupings are highlighted in gray, e.g., SP, OL, and II agree on average on a third of edges. Combi refers to intersection combination of all four methods, SP to Sign Pattern, OL to Overlap, II to Interaction Information, and DPI to Data Processing Inequality. Less prominent groupings are aggregated with others.

Statistic	Minimum	1st Quartile	Median	Mean	2nd Quartile	Maximum
SP	0	0	0.2	0.3	0.5	3.7
OL	0	0	0.1	0.2	0.3	2.0
II	0	0.7	1.3	1.4	2.0	6.0
DPI	0	0.1	0.3	0.4	0.6	2.6
SP and OL	4.9	12.2	14.9	15.0	17.5	30.0
SP, OL, and II	19.1	29.5	32.6	32.8	36.2	49.6
SP, OL, and DPI	2.6	7.1	8.9	9.1	10.8	22.1
SP, OL, II, DPI, and Combi	22.4	32.1	35.6	35.5	38.6	48.6
other	0.4	3.3	4.9	5.1	6.6	15.4

Supplementary Table S2: **Performance of environmentally-driven edge detection methods on simulated networks.** These include 50 microorganisms and 1225 possible associations. Values display median (standard deviation) for simulated networks and simulated networks incorporating noise. Combi refers to intersection combination of all four methods, SP to Sign Pattern, OL to Overlap, II to Interaction Information, and DPI to Data Processing Inequality. The methods with highest (TP, TN, TPR, TNR, PPV, ACC) or lowest (FP, FN, FPR) median, respectively, are highlighted in gray.

Method	Combi	SP	OL	II	DPI
without noise					
number of nodes	50 (0.045)	47 (6.6)	48 (5.6)	50 (0.94)	50 (0.1)
number of edges	737 (50)	140 (52)	144 (58)	354 (67)	601 (60)
TP	332 (47)	893 (64)	888 (69)	696 (72)	459 (53)
TN	45 (5.1)	8 (4.3)	9 (4.7)	24 (5.8)	37 (5.5)
FP	15 (4.6)	51 (5.8)	51 (6.2)	36 (6.4)	23 (5.2)
FN	692 (48)	131 (49)	136 (54)	330 (63)	564 (56)
TPR	0.32 (0.04)	0.87 (0.05)	0.87 (0.05)	0.68 (0.06)	0.45 (0.05)
TNR	0.75 (0.07)	0.14 (0.07)	0.15 (0.08)	0.4 (0.10)	0.62 (0.08)
FPR	0.25 (0.07)	0.86 (0.07)	0.85 (0.08)	0.6 (0.10)	0.38 (0.08)
PPV	0.96 (0.011)	0.95 (0.005)	0.95 (0.005)	0.95 (0.007)	0.95 (0.009)
ACC	0.35 (0.04)	0.83 (0.04)	0.83 (0.048)	0.66 (0.057)	0.46 (0.046)
with noise					
number of nodes	50 (0.08)	47 (5.6)	48 (4.9)	50 (0.47)	50 (0.12)
number of edges	828 (56)	144 (53)	149 (59)	428 (79)	717 (73)
TP	219 (48)	864 (69)	860 (72)	605 (81)	324 (64)
TN	49 (5)	9 (4.6)	9 (4.9)	29 (6.3)	42 (5.8)
FP	10 (3.9)	50 (6.1)	50 (6.4)	30 (6.6)	17 (5.1)
FN	779 (53)	137 (50)	139 (55)	398 (75)	674 (69)
TPR	0.22 (0.05)	0.86 (0.05)	0.86 (0.06)	0.6 (0.08)	0.32 (0.06)
TNR	0.84 (0.07)	0.15 (0.08)	0.16 (0.08)	0.49 (0.1)	0.72 (0.09)
FPR	0.16 (0.07)	0.85 (0.08)	0.84 (0.08)	0.51 (0.1)	0.28 (0.09)
PPV	0.96 (0.014)	0.95 (0.005)	0.95 (0.005)	0.95 (0.007)	0.95 (0.012)
ACC	0.25 (0.04)	0.82 (0.05)	0.82 (0.05)	0.6 (0.07)	0.34 (0.06)

SP - Sign Pattern; OL - Overlap; II - Interaction Information; DPI - Data Processing Inequality; Combi-intersection combination

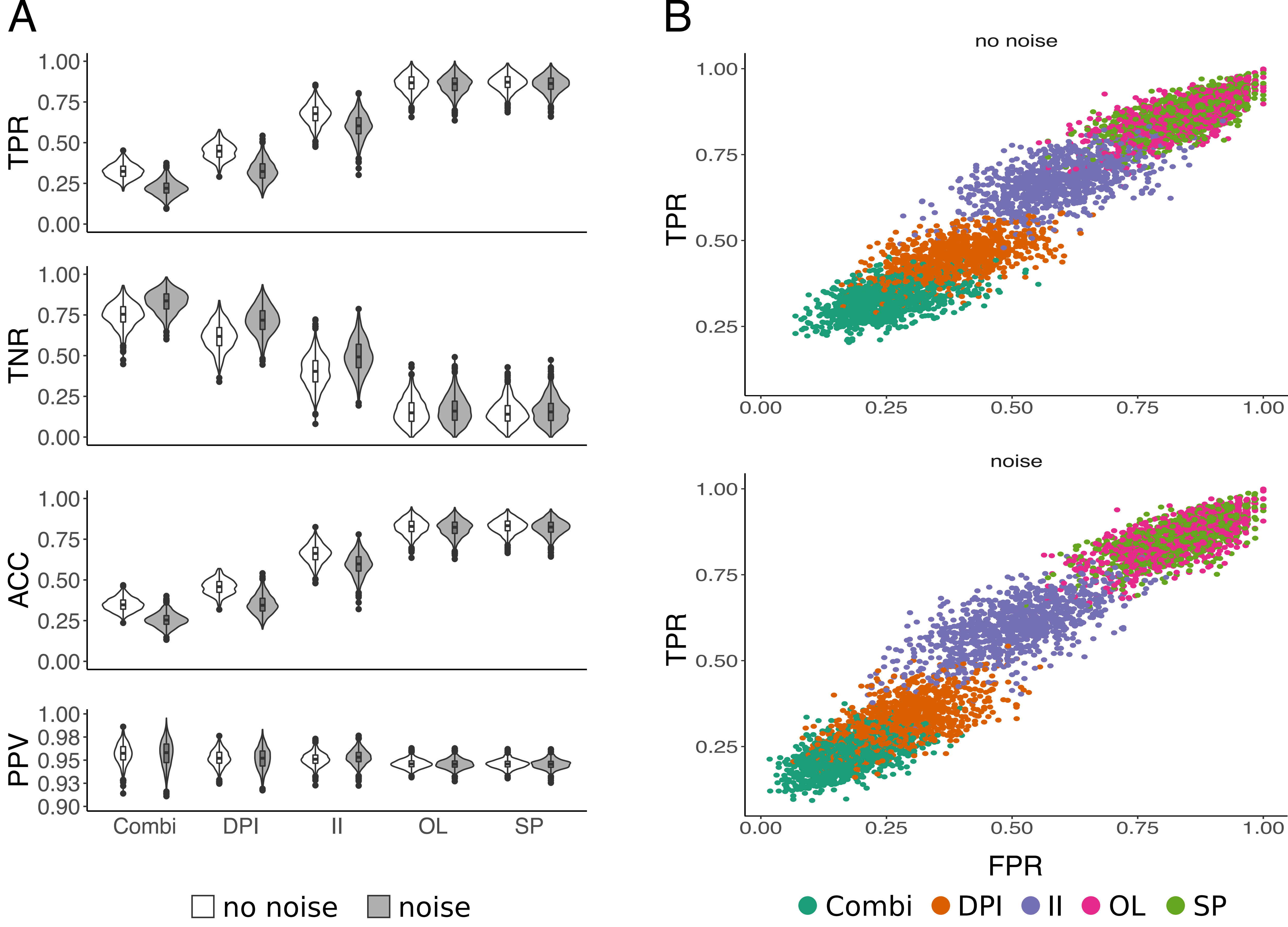
Supplementary Table S3: **Number of triplets a microbial edge is part of in the BBMO network.** SP and OL not listed below because they remove 100% of microbial associations that are within at least one triplet. The total number of edges (all) is given along the number of positive (pos) and negative (neg) edges. Combi refers to intersection combination of all four methods, II to Interaction Information, and DPI to Data Processing Inequality.

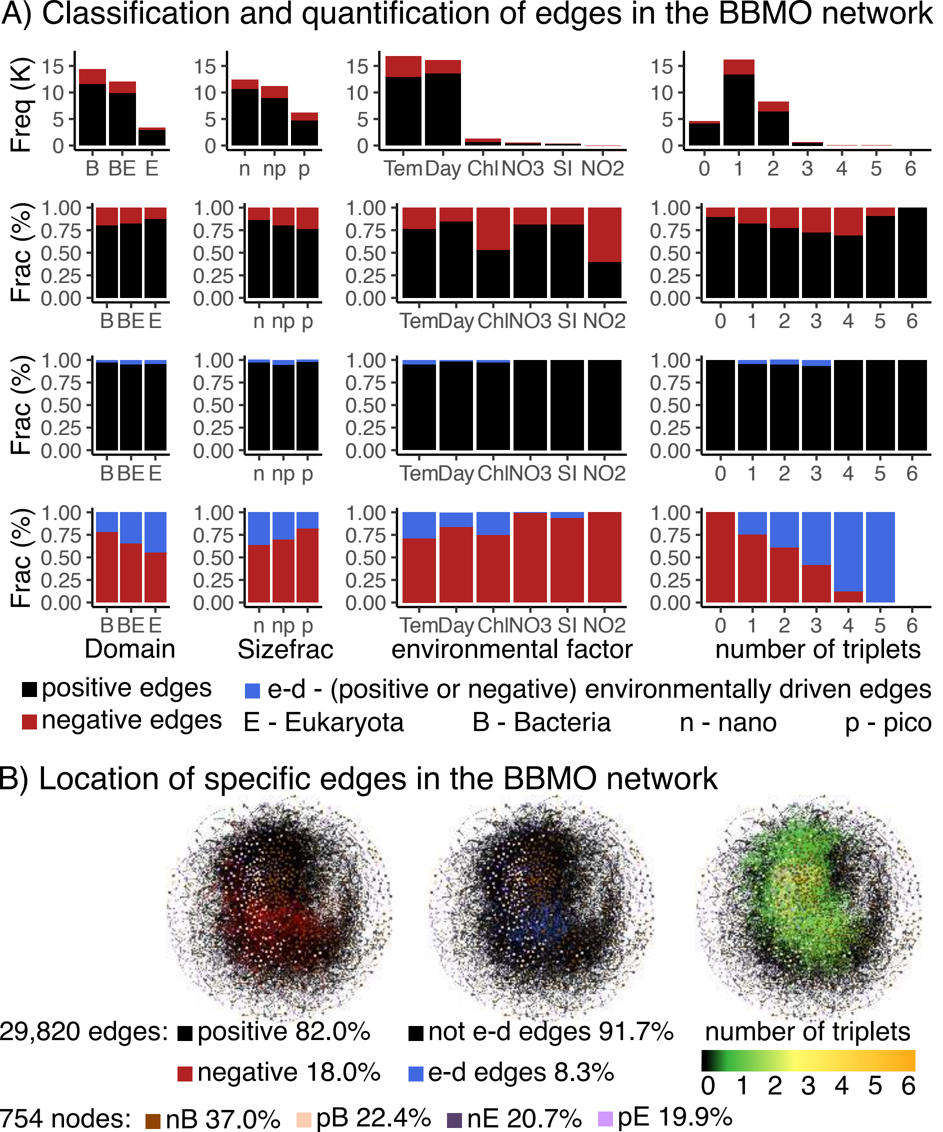
Triplets	all	pos (%)	neg (%)	Combi (%)	II (%)	DPI (%)
0	4 590	4 124 (89.8)	466 (10.2)	NA	NA	NA
1	16 193	13 369 (82.6)	2 824 (17.4)	1 276 (7.9)	3 851 (23.8)	4 560 (28.2)
2	8 266	6 404 (77.5)	1 862 (22.5)	1 048 (12.7)	3 335 (40.3)	2 585 (31.3)
3	667	484 (72.6)	183 (27.4)	140 (21.0)	388 (58.2)	222 (33.3)
4	81	56 (69.1)	25 (30.9)	22 (27.2)	75 (92.6)	25 (30.9)
5	22	20 (90.9)	2 (9.1)	2 (9.1)	22 (100)	2 (9.1)
6	1	1 (100)	NA	NA	1 (100)	NA

Supplementary Table S4: **The BBMO network based on real data.** The BBMO network contained bacteria (B) and eukaryotes (E) from the picoplankton (p) and nanoplankton (n). This table summarizes the number and fraction of microbial associations classified by EnDED as environmentally-driven. Combi refers to the intersection combination of all four methods, II to Interaction Information, and DPI to Data Processing Inequality. Both methods, Sign Pattern and Overlap, are not shown because both remove all microbial edges found in at least one triplet. For example (last row), 349 (14.9%) associations between bacteria from the picoplankton with eukaryotes from the nanoplankton were classified by intersection combination as environmentally-driven (indirect), II classified 30.6% and DPI 37.2% as environmentally-driven.

Type	edges	positive	negative	triplets	Combi	II	DPI
nB	6 377	5 453 (85.5)	924 (14.5)	5 150 (80.8)	376 (5.9)	1 512 (23.7)	1 080 (16.9)
n+pB	5 191	4 069 (78.4)	1 122 (21.6)	4 824 (92.9)	440 (8.5)	1 381 (26.6)	1 678 (32.3)
pB	2 832	2 053 (72.5)	779 (27.5)	2 160 (76.3)	125 (4.4)	569 (20.1)	631 (22.3)
nE	1 319	1 163 (88.2)	156 (11.8)	1 016 (77.0)	113 (8.6)	350 (26.5)	254 (19.3)
n+pE	1 165	976 (83.8)	189 (16.2)	1 006 (86.4)	158 (13.6)	353 (30.3)	370 (31.8)
pE	895	820 (91.6)	75 (8.4)	543 (60.7)	44 (4.9)	153 (17.1)	113 (12.6)
nB+E	4 703	4 080 (86.8)	623 (13.2)	4 120 (87.6)	438 (9.3)	1 345 (28.6)	1 043 (22.2)
pB+E	2 520	1 908 (75.7)	612 (24.3)	1 980 (78.6)	204 (8.1)	626 (24.8)	647 (25.7)
nB+pE	2 483	2 100 (84.6)	383 (15.4)	2 222 (89.5)	241 (9.7)	668 (26.9)	709 (28.6)
pB+nE	2 335	1 836 (78.6)	499 (21.4)	2 209 (94.6)	349 (14.9)	715 (30.6)	869 (37.2)

B - Bacteria; E - Eukaryotes; n - nano fraction; p - pico fraction


A
TPR
TNR
ACC
PPV
1.00
0.75
0.50
0.25
0.00
0.98
0.95
0.93
0.90
Combi
DPI
II
OL
SP
□ no noise
▨ noise
B
no noise
noise
TPR
FPR
0.00
0.25
0.50
0.75
1.00
● Combi ● DPI ● II ● OL ● SP



A) Classification and quantification of edges in the BBMO network

B) Location of specific edges in the BBMO network

environment f
+/-
+/-
+/-
microbe v
microbe w
keep association
remove association

Sign Pattern

+++, +--, -+-, --+

vs

---, -++, +-+, ++-

Overlap in time

vs

Interaction Information

II(v,w,f) < 0

vs

II(v,w,f) > 0

Data Processing Inequality

MI(v;w) < MI(v;f) and
MI(v;w) < MI(w;f)

vs

MI(v;w) > MI(v;f) or
MI(v;w) > MI(w;f)

Entropy

S(v) = $-\sum_{i=1}^{n} p(v_i)\log\left(p(v_i)\right)$

Mutual Information

MI(v;w) = S(v) + S(w) - S(v,w)

Conditional Mutual Information

CMI(v;w|f) = S(v,f) + S(w,f) - S(v,w,f) - S(f)

Interaction Information

II(v,w,f) = CMI(v;w|f) - MI(v;w)