Disentangling ecological networks in marine microbes

Ina Maria Deutschmann

Director of thesis

Dr. Ramiro Logares, Institut de Ciències del Mar (ICM-CSIC)

Tutor

Prof. Oriol Serra, Universitat Politècnica de Catalunya (UPC)

School of Mathematics and Statistics
Barcelona, 15. July 2021

für meine Eltern Cornelia und Jürgen Deutschmann

Abstract

There is a myriad of microorganisms on Earth contributing to global biogeochemical cycles, and their interactions are considered pivotal for ecosystem function. Previous studies have already determined relationships between a limited number of microorganisms. Yet, we still need to understand a large number of interactions to increase our knowledge of complex microbiomes. This is challenging because of the vast number of possible interactions. Thus, microbial interactions still remain barely known to date. Networks are a great tool to handle the vast number of microorganisms and their connections, explore potential microbial interactions, and elucidate patterns of microbial ecosystems.

This thesis locates at the intersection of network inference and network analysis. The presented methodology aims to support and advance marine microbial investigations by reducing noise and elucidating patterns in inferred association networks for subsequent biological down-stream analyses. This thesis's main contribution to marine microbial interactions studies is the development of the program EnDED (**En**vironmentally-**D**riven **E**dge **D**etection), a computational framework to identify environmentally-driven associations inside microbial association networks, inferred from omics datasets. We applied the methodology to a model marine microbial ecosystem at the Blanes Bay Microbial Observatory (BBMO) in the North-Western Mediterranean Sea (ten years of monthly sampling). We also applied the methodology to a dataset compilation covering six global-ocean regions from the surface (3 m) to the deep ocean (down to 4539 m). Thus, our methodology provided a step towards studying the marine microbial temporal patterns and the distribution in space via the horizontal (ocean regions) and vertical (water column) axes.

To reach accurate interaction hypotheses, it is important to determine, quantify, and remove environmentally-driven associations from marine microbial association networks. Moreover, our results underlined the need to study the dynamic nature of networks, in contrast to using single static networks aggregated over time or space. Our novel methodologies can be used by a wide array of researchers investigating networks and interactions in diverse microbiomes.

Resumen

Hay una gran cantidad de microorganismos en la Tierra que contribuyen a los ciclos biogeoquímicos globales, y sus interacciones se consideran fundamentales para la función del ecosistema. Estudios previos ya han determinado relaciones entre un número limitado de microorganismos. Sin embargo, todavía necesitamos comprender una gran cantidad de interacciones para aumentar nuestro conocimiento de los microbiomas más complejos. Esto representa un gran desafío debido a la gran cantidad de posibles interacciones. Por lo tanto, las interacciones microbianas son aun poco conocidas. Las redes representan una gran herramienta para analizar la gran cantidad de microorganismos y sus conexiones, explorar posibles interacciones y dilucidar patrones en ecosistemas microbianos.

Esta tesis se ubica en la intersección entre la inferencia de redes y el análisis de redes. La metodología presentada tiene como objetivo avanzar las investigaciones sobre interacciones microbianas marinas mediante la reducción del ruido en las inferencias de redes y elucidar patrones en redes de asociación permitiendo análisis biológicos posteriores. La principal contribución de esta tesis a los estudios de interacciones microbianas marinas es el desarrollo del programa EnDED (**En**vironmentally-**D**riven **E**dge **D**etection), un marco computacional para identificar asociaciones generadas por el medio ambiente en redes de asociaciones microbianas, inferidas a partir de datos ómicos. Aplicamos la metodología a un modelo de ecosistema microbiano marino en el Observatorio Microbiano de la Bahía de Blanes (BBMO) en el Mar Mediterráneo Noroccidental (diez años de muestreo mensual). También, aplicamos la metodología a una compilación de conjuntos de datos que cubren seis regiones oceánicas globales desde la superficie (3 m) hasta las profundidades del océano (hasta 4539 m). Por lo tanto, nuestra metodología significa un paso adelante hacia de los patrones temporales microbianos marinos y el estudio de la distribución microbiana marina en el espacio a través de los ejes horizontal (regiones oceánicas) y vertical (columna de agua).

Para llegar a hipótesis de interacción precisas, es importante determinar, cuantificar y eliminar las asociaciones generadas por el medio ambiente en las redes de asociaciones microbianas marinas. Además, nuestros resultados subrayaron la necesidad de estudiar la naturaleza dinámica de las redes, en contraste con el uso de redes estáticas únicas agregadas en el tiempo o el espacio. Nuestras nuevas metodologías pueden ser utilizadas por una amplia gama de investigadores que investigan redes e interacciones en diversos microbiomas.

Resum

Hi ha una infinitat de microorganismes a la Terra que contribueixen als cicles biogeoquímics mundials i les seves interaccions es consideren fonamentals pel funcionament dels ecosistemes. Estudis previs ja han determinat les relacions entre un nombre limitat de microorganismes. Tot i això, encara hem d'entendre un gran nombre d'interaccions per augmentar el nostre coneixement dels microbiomes complexos. Això és un repte a causa del gran nombre d'interaccions possibles. Per això, les interaccions microbianes encara són poc conegudes fins ara. Les xarxes són una gran eina per tractar el gran nombre de microorganismes i les seves connexions, explorar interaccions microbianes potencials i dilucidar patrons d'ecosistemes microbians.

Aquesta tesi es sitúa a la intersecció de la inferència de xarxes i l'anàlisi de la xarxes. La metodologia presentada té com a objectiu donar suport i avançar en investigacions microbianes marines reduïnt el soroll i dilucidant patrons en xarxes d'associació inferides per a posteriors anàlisis biològiques. La principal contribució d'aquesta tesi als estudis d'interaccions microbianes marines és el desenvolupament del programa EnDED (**En**vironmentally-**D**riven **E**dge **D**etection), un marc computacional per identificar associacions impulsades pel medi ambient dins de xarxes d'associació microbiana, inferides a partir de conjunts de dades òmics. Vam aplicar la metodologia a un model d'ecosistema microbià marí a l'Observatori Microbià de la Badia de Blanes (BBMO) al mar Mediterrani nord-occidental (deu anys de mostreig mensual). També hem aplicat la metodologia a una recopilació de dades que cobreix sis regions oceàniques globals des de la superfície (3 m) fins a l'oceà profund (fins a 4539 m). Per tant, la nostra metodologia va proporcionar un pas cap a l'estudi dels patrons temporals microbians marins i la distribució microbiana marina a l'espai a través dels eixos horitzontal (regions oceàniques) i vertical (columna d'aigua).

Per arribar a hipòtesis d'interacció precises, és important determinar, quantificar i eliminar associacions impulsades pel medi ambient de les xarxes d'associació microbiana marina. A més, els nostres resultats van subratllar la necessitat d'estudiar la naturalesa dinàmica de les xarxes, en contrast amb l'ús de xarxes estàtiques individuals agregades al llarg del temps o l'espai. Les nostres noves metodologies poden ser utilitzades per una àmplia gamma d'investigadors que investiguen xarxes i interaccions en diversos microbiomes.

Acknowledgments

My journey was full of diverse challenges, enriching experiences, and amazing colleagues, mentors, friends, and family. I am thankful for their company during the process and progress.

With sincere gratitude, I would like to acknowledge the support of the project supervisor, Dr. Ramiro Logares. My journey started with his project at the Marine Science Institute (ICM-CSIC) in the *Ecology of Marine Microbes* group as part of the EU H2020 Marie Skłodowska-Curie Innovative Training Network project SINGEK (Coordinator: Dr. Ramon Massana). Thank you for creating the possibility for this project, your guidance, detailed and valuable manuscript revisions. A big thanks to the ThePaperMill team (Dr. Gavin Lucas, Dr. Tobias Maier) for the unconventional and informative transferable skills workshops, including a follow-up writing service – Thank you, Dr. Valeria Di Giacomo. I dedicate a specific Thanks to the SINGEKs project manager Dr. Elena Torecilla. You supported me in diverse matters, especially when I faced the Spanish/Catalan language barriers. I would also like to thank Dr. Ramiro Logares, Dr. Ramon Massana, Prof. Josep M. Gasol, and Dr. Silvia González Acinas for providing me access to the high-grade datasets (BBMO, Malaspina, and Hotmix), which are very valuable within the field of marine microbial science. In addition, you, Dr. Celia Marrasè, Dr. Sergio M. Vallina, Vanessa Balagué, Dr. Caterina R. Giner, Dr. Marta Sebastián, and Prof. Carlos M. Duarte provided valuable feedback, clarifications, ecological-driven suggestions, and further inspirations. Thank you!

I want to express my deep gratitude to my hosts during the research visits to the *Microbial Systems Biology* lab in Leuven, Belgium (hosts: Prof. Karoline Faust and Dr. Gipsi Lima-Mendez) and the *Computational Biology* group in Nantes, France (hosts: Prof. Damien Eveillard and Dr. Samuel Chaffron). Your labs provided excellent research conditions and a diverse network-oriented group of scientists enabling fruitful conversations that advanced my ideas, approaches, and, subsequently, projects. I am grateful for the discussions that influenced my work tremendously.

After becoming a Master of Biomathematics, the next step in my career development is the Ph.D. in Applied Mathematics. I feel fortunate to have been accepted to the Mathematics and Statistics school at UPC. I want to thank my tutor, Prof. Oriol Serra, as well as Carme Capdevilla (administration) and Juanjo Rue (doctoral program coordinator). Your indispensable assistance to the administrative work and clarifications on diverse regulations is greatly appreciated.

Setting the stage for a diverse audience was demanding. It necessitated setting the ground for and picking up mathematicians, computational biologist, and marine microbial ecologist. For that, I am thankful for the valuable feedback and proofreading on preliminary and/or advanced sections from Dr. Marko Budinich, Claudia Kohring, Carlota Ruiz Gazulla, Dr. Anders K. Krabberød, Dr. Georgina Brennan, Dr. Nils Giordano, Adam Mitchinson, Dr. Carolin Malsch, Ole Geldschläger, and Fran Latorre. Moreover, I am grateful for proofreading the additional analyses (Chapter 8) and/or sections in the discussion to Dr. Marko Budinich, Dr. Nils Giordano, Dr. Anders K. Krabberød, and Dr. Georgina Brennan.

Diverse people supported, encouraged, and motivated me. Thank you, my SINGEK companions, Francois, Felipe, Ati, Imer, Jari, Laura, and Alex, for your open ears and encouragement. I want to thank my ICM office mates from day one, Manu, Dani, and Paula, as well as Carlota, Clara, and Pedro. You accompanied and motivated me on my journey of oscillating frustrations and wins. I want to thank my 2019-Nantes office buddy Marko. Your expertise and openness to discuss research lead to a rapid advance in my projects despite the little time I could be there. Thank you for your kind and supportive way of being. I look happily back on my time in Nantes thanks to you, Sam, Damien, Erwan, Nils, Benjamin, and Johanna. I want to thank my former colleagues from the Master's studies Carolin, Ole, and Elisa. You openly gave your support when I faced mathematical difficulties or needed to discuss mathematical ideas. Thank you, Claudia K., Chenna, Adam, and Erwan, for your support and curiosity in my work.

I am thankful for the many more magnificent people who supported and encouraged me from near and far. Thank you, Alexandrine and Ann-Christin, my POD#13 crew (especially Deb, Russ, Lon, and Patrick), Charlie, Claudia, Ari, Andreu, Miguel, Guillem, Jananan, Fran, Lidia, Aleix, Pablo, Chrissi, Vivien, Stephan, Diana, Sylvia, and especially Marie, Kai, Anne, my parents and grandparents. Being the first in my family that goes the academic path, my journey appears very foreign. I am indebted to my family for their unconditional support and patience. Despite the many challenges I faced during my doctoral time, I knew they were insignificant compared to the quest of my dear friend Giovanni. I am immensely grateful for your endurance, enabling you to cheer me up and motivating me from the other side of the big ocean.

Danke! Thank you! Gracias! Gràcies!
Bedankt! Merci! Takk! Grazie! Obrigado! Hvala! ధన్యవాదాలు! *با تشکر!*

Contents

List of figures

List of tables

General introduction

A graph is a simple mathematical object and can be found everywhere.

The mathematical field of *Graph Theory* started with a riddle that entertained people in Königsberg in 1736 (Biggs *et al.*, 1986; Barabási, 2003), when Leonhard Euler translated the real-world problem into an abstract mathematical problem (Euler, 1741). *Graph Theory* became an export hit in various scientific fields. In general, graphs capture relationships between entities or objects, and they are the mathematical way to represent networks (Koutrouli *et al.*, 2020). In turn, networks are powerful tools to represent complex systems (Barabási, 2003; Amaral & Ottino, 2004). Thus, not surprisingly, they are found *everywhere* as they have diverse applications (Estrada & Knight, 2015), for example in biology.

Various biological networks have been studied in the past decades and, for years to come, graph-theoretic methods will remain indispensable tools to further understand complex interconnected systems (Alm & Arkin, 2003; Faust & Raes, 2012; Layeghifard *et al.*, 2017; Röttjers & Faust, 2018; Sporns, 2018). Alm and Arkin give two partial answers to what we can learn about biology by studying networks (Alm & Arkin, 2003): i) the use of network-based approaches to uncover patterns help to organize the vast collections of data, making them more accessible to and valuable for biologists; ii) the reformulation of existing biological questions from a network perspective has the potential to include all available data and to answer otherwise unsolvable questions.

Although their applications and functions differ tremendously, real-world networks share universal properties. Networks have been characterized in the late nineties: networks are a *small world* (Watts & Strogatz, 1998), and the vast majority of networks are *scale-free* (Barabási & Albert, 1999). Thus, we can find a beautiful universality in network architecture. Moreover, graph-theoretic concepts applied in one field can be applied in another, e.g., studying the tiniest living beings on Earth, the micrrobes (microorganisms).

A microorganism is the smallest (simplest) living entity and can be found everywhere.

All living beings developed from a single microbial cell, and life without microorganisms would not be possible. Our human body provides a permanent albeit dynamic home to a vast number of microorganisms. They live within and outside us. The human microbiome, i.e., the set of microorganisms that colonize humans, is composed of about 3.8×10^{13} bacterial cells exceeding the number of human eukaryotic cells by a factor of 1.3 (Sender *et al.*, 2016). The largest portion of the human microbiome is located in the digestive tract; the gut microbiota of a 70-kg person would weigh 0.2 kg (Sender *et al.*, 2016).

While these numbers are already impressive, microorganisms play a far more significant role for the Earth. They are considered fundamental for the functioning of the global ecosystem (DeLong, 2009; Krabberød *et al.*, 2017), and for the ocean biogeochemical cycling (Falkowski *et al.*, 2008). The estimated number of microorganisms is around 10^{30} cells (Whitman *et al.*, 1998; Kallmeyer *et al.*, 2012). They constitute 60% of the biomass on Earth and carry out more photosynthesis than green plants (Demain & Adrio, 2008). In the sunlit ocean, photosynthetic microorganisms are responsible for ~50% of carbon fixation on Earth (Field *et al.*, 1998). In the marine environment, microorganisms account for about 70% of the total marine biomass (Bar-On *et al.*, 2018).

Although microorganisms dominate our world, they are the smallest living entities on Earth. Therefore, only the technological advance of the past centuries allowed their detection and quantification. To understand the microbial ecosystem, we need to know which microorganisms are there, how many microorganisms are there, and what influences them (environmental dependence). Extensive research elucidated these aspects. However, we also need to know how microorganisms are connected: we need to know who is interacting with whom. Despite the tremendous importance of microorganisms, their interactions are still barely known (Krabberød *et al.*, 2017; Bjorbækmo *et al.*, 2019).

A network is a perfect tool to model interactions between microorganisms.

The vast microbial diversity and the fact that most microorganisms are still uncultured (Baldauf, 2008; Lewis *et al.*, 2020) make it impossible to experimentally test all potential interactions. However, omics-technologies allow to estimate microbial sequence abundances over spatial and temporal scales and determine interaction hypotheses, e.g., via association analysis (Röttjers & Faust, 2018). These associations constitute a network, with nodes and edges representing microorganisms and potential interactions, respectively (Weiss *et al.*, 2016; Layeghifard *et al.*, 2017). Networks' inference and biological interpretation are in their infancy (Lv *et al.*, 2019) with remaining open challenges (Faust, 2021). Nevertheless, as microorganisms are highly interconnected (Layeghifard *et al.*, 2017), association networks provide a general overview of the entire microbial system and are valuable for generating interaction hypotheses.

The awareness of the importance of microorganisms within the ecosystem has increased during the last decades. In particular, it has benefited from the development of new omics tools (e.g., genomics and metagenomics), analytical methods (i.e., algorithms), and computing performance. Therefore, currently, it is possible to explore microbial diversity, species distribution, and metabolic function in more depth. Such advances are having an essential impact on microbial ecology and microbiology. In particular, microbiologists are changing their perspective from a classical reductionist one (concentrating on one microbial group, species, or

metabolism), to another aiming to understand the functioning of ecosystems. Analyzing and converting microbiome data into meaningful biological insights is challenging (Layeghifard *et al.*, 2017). Layeghifard *et al.* (2017) state that for understanding microbial ecosystems, it is essential to understand microbial interactions. Subsequently, to understand marine microbial interactions, it is essential to provide marine microbial ecologist with a clearer picture of true interactions, i.e., more truthful marine microbial associations.

Right now, we live in a time in which collaborations between different research fields are becoming more and more necessary. For instance, ecologists profit from other specialists trained in algorithms and programming to investigate their data and extract information that ecologists are able to use for their investigations. In exchange, specialists (e.g., an applied- or bio-mathematician), can profit from applying their theories on data from the natural world. Researchers trained in mathematics, computer science, and with a background in biology are required to make use of the technological advances in omics. Specifically, for this Ph.D. project, these advances allowed us to investigate marine microbial interactions at a great depth. This thesis is crossing fields between molecular biology, ecology, omics technologies, mathematics, especially network theory, and computer science.

Previous network-based investigations have contributed to our understanding of marine microbial interactions. Still, what we know is only a drop in the ocean, in comparison to things yet to be discovered. Networks are a great tool to handle the vast number of microorganisms and their connections, explore potential microbial interactions, and elucidate patterns of microbial ecosystems. However, diverse challenges exist (see Chapter 4, and a recent listing of ten underexplored challenges (Faust, 2021)). Faust (2021) recently concluded that, if we want to learn more from microbial networks, we need to broaden our focus of research beyond inference algorithms and tackle underexplored challenges for microbial network construction and analysis. This thesis is located between network inference and interpretation. The main aims are presented in the next section.

Aims and Outline

Aims

The main aims of this thesis are

- to improve the prediction of association networks for biological downstream analysis;
- to provide better interaction hypotheses;
- to elucidate temporal patterns;
- to elucidate spatial patterns.

The path from an initial constructed network to an interpretable one presents various challenges, as summarized in Chapter 4 that give context to this thesis. It would be challenging to experimentally tackle this matter, considering the vast number of microorganisms and potential interactions. Thus, networks are a great tool to generate interaction hypotheses, i.e., a set of potential interactions. Reducing the number of interaction hypotheses reduces the potential interactions to be experimentally tested.

Objectives

The project is located at the intersection of network inference and biological interpretation. In each subproject, we proposed steps that should be done after the construction and before the interpretation of a network to improve it before biological downstream analysis. Consequently, the following objectives had driven the doctoral subprojects:

- Disentangle environmental effects: environmentally-driven association due to the environmental preferences of microorganisms or true interaction?
- Unravel the temporal nature: permanent, seasonal, or temporal association?
- Elucidate temporal patterns in a network constructed from a model marine microbial ecosystem at the Blanes Bay Microbial Observatory (BBMO) in the North-Western Mediterranean Sea.
- Resolve the spatial distribution: global or regional association?
- Enlighten depth-specific patterns in a network constructed from a compilation of four datasets covering different depths in global open oceans (North and South Atlantic, North and South Pacific, and Indian Oceans), and Mediterranean Sea.
- Improve interaction hypotheses: select the associations that most likely are true interactions and reduce the set to the most interesting candidates for experimental testing by identifying the associations with highest temporal and spatial recurrence, respectively.

Outline

In Part I, we introduced biological and graph-theoretic aspects. Moreover, we described association networks to study microbial interactions and the associated challenges. Results were then presented in Part II, which is a collection of three (submitted) manuscripts representing the main doctoral subprojects, and additional analyses. Finally, in Part III, we presented further discussions, future perspectives, and conclusions.

Methodology

The thesis aimed to increase our knowledge of microbial interactions in the oceans and the Mediterranean Sea by using omics approaches and tools from graph theory (networks). By combining them, it is possible to generate hypotheses on microbial interactions. Specifically, we investigated putative ecological interactions via marine microbial association networks. We used existing network construction tools on microbial data (16S and 18S rRNA gene data) either originating from a marine microbial observatory (temporal data), or from the global ocean (spatial data).

Results

The work provided further puzzle pieces to the quest on elucidating the microbial world. It widened the frontiers in marine microbial ecology and this was a step forward to better understand marine microbial ecosystems. More precisely, this thesis increased our knowledge on:

- Associations between microbial taxa;
- The architecture of networks displayed by ocean microorganisms;
- How the network changes in time, with depth, and between oceans.

Specifically, we provided the following improvements, which are further described in the three main doctoral subprojects:

- a program to filter out environmental influence;
- a methodology to generate a temporal network allowing to investigate temporal patterns over ten years for a model marine microbial ecosystem at the Blanes Bay Microbial Observatory in the North-Western Mediterranean Sea;
- a methodology to generate sample-specific subnetworks allowing to explore horizontal and vertical patterns along spatial scales. It uses a compilation of four datasets covering five ocean basins (North and South Atlantic, North and South Pacific, and Indian Oceans), and the Mediterranean Sea from the surface to the deep ocean.

In chapter 5, we aimed to distinguish true ecological interactions from associations representing environmental preference. This solved the problem of indirect dependencies due to environmental factors. We then applied our method to simulated data and an association network of a model marine microbial ecosystem at the Blanes Bay Microbial Observatory (BBMO) in the North-Western Mediterranean Sea. We managed to reduce the number of edges, indirectly tackling the problem of dense networks. Moreover, we obtained a smaller and stronger set of potential associations, simplifying the testing of interaction hypotheses. However, there were still too many potential associations. Thus, we proposed another step of measuring association recurrence since an association with high recurrence (temporal and spatial) may be more likely to represent a true interaction.

In chapter 6, we aimed to unravel the temporal nature and pattern of associations. This project took the marine microbial association studies on the model marine microbial ecosystem at the BBMO a step forward, by introducing the temporal dimension. We used the single static network and microbial sequence abundance to determine sample-specific (monthly) subnetworks that constituted a temporal network, i.e., each subnetwork was a layer in the temporal network. These subnetworks potentially represent a valid temporal game-changer in the field of marine microbial association studies until new data will allow to construct sample-specific networks (note: *networks*, not *subnetworks*). The temporal network allowed to distinguish between permanent, seasonal, and temporary associations. Hereby, we pinpointed the properties of season-specific global networks.

In chapter 7, building upon strategies from the former two, we disentangled the spatial distribution of associations in a marine microbial network of the global ocean and Mediterranean Sea. Using a compilation of datasets, we constructed a network covering different locations and depths. We adapted the sample-specific subnetworks approach, and aimed to distinguish between associations that are endemic to certain regions and global ones. Moreover, we determined the change of the microbial association networks between different depths and regions. Hereby, we pinpointed the properties of depth-specific subnetworks and identified clusters of similar subnetworks.

Part I Background

Chapter 1 Biological aspects

The essential role of the ocean microbiome

Marine microorganisms have critical functions in ecosystems. They contribute directly or indirectly to shaping and maintaining our current environment (Falkowski *et al.*, 2008). In the ocean, they are essential players in biogeochemical cycles[1] (DeLong, 2009). In particular, the smallest ocean microorganisms (so-called picoplankton) have a crucial role in the global carbon cycle (Worden *et al.*, 2015). They account for a significant fraction of the total atmospheric carbon fixation in the ocean (Li, 1994; Jardillier *et al.*, 2010; Massana, 2011) and for about 50% of the global primary productivity (Field *et al.*, 1998). Besides the primary producer, heterotrophic marine picoplankton that preys on other picoplankton has a fundamental role in respiring organic carbon and channeling it to upper trophic levels when being preyed on itself (del Giorgio & Duarte, 2002; Massana, 2011). The ocean picoplankton constitutes the base of the marine food web and, subsequently, the marine ecosystem (Worden *et al.*, 2015).

Two types of microorganisms populate the ocean picoplankton: prokaryotes (including bacteria and archaea), and small eukaryotes. They feature fundamental differences in cellular structure, feeding habits, diversity of metabolism, growth rates, and ecological behavior (Massana & Logares, 2013; Keeling & Campo, 2017). Microbial communities are not a mere collection of independent individuals; they are interconnected ecological entities that communicate, cross-feed, recombine, and co-evolve (Layeghifard *et al.*, 2017). Thus, to understand microbial ecosystems, it is essential to understand microbial interactions (Layeghifard *et al.*, 2017).

Marine microbial interactions

Any type of biological interaction between two individual organisms is called *symbiosis*[2]. The organisms, so-called *symbionts*, either belong to the same or different species. Ecological interactions between pairs of symbionts can be positive (beneficial), neutral, or negative (disadvantageous), with consequences for one or both symbionts. The different types of symbiosis are classified according to their ecological effect on the organisms (Faust & Raes, 2012) (Table 1).

[1] Biogeochemical cycle, also called cycling of substances, is a pathway in which a chemical substance (e.g., carbon, oxygen, or nitrogen) moves through biotic and abiotic parts of the Earth. In some systems, such as an ocean, there are reservoirs where a substance remains for a long period of time. Ecological systems have many biochemical cycles that function as a part of them. All substances (chemical elements) that are present in organisms are part of biochemical cycles.

[2] Traditionally, symbiosis referred to mutualistic relationships. Here, symbiosis is used in the broader sense to include all interactions. Martin & Schwab (2013) present a survey of the usage of the term symbiosis.

Table 1: **Ecological interactions** are categorized on the basis of the pairwise combination of the consequences for the two individual microorganisms. Such consequences are either beneficial (+), neutral (0), or disadvantageous (-). Mutualism: both interaction partner benefit. Commensalism: one benefits while the other neither has an advantage nor a disadvantage. Antagonism: one benefits while the other has a disadvantage (win-loss interactions, e.g., parasitism and predation). Amensalism: one has a disadvantage while the interaction is neutral for the other. Competition: both have a disadvantage. Neutralism: both are neither positively nor negatively impacted.

Ecological interaction	The consequence for species 1	The consequence for species 2
Mutualism	+	+
Commensalism	+	0
Antagonism	+	-
Amensalism	0	-
Competition	-	-
Neutralism	0	0

Microbial interactions may involve physical contact between the two partners or not. Within microbial communities, non-physical interactions can be substantial. For example, one microorganism releases toxic chemicals that inhibit the growth of others, or produces and releases a substance that may be essential for other members of the community. Thus, microbial interactions can also be predicted from a metabolic perspective. For example, complementarity in metabolite requirement and production in different microorganisms point towards an interaction (Borenstein & Feldman, 2009). Borenstein & Feldman (2009) introduced a pairwise measure that reflects the extent to which the nutritional requirements of one species could be satisfied by the biosynthetic[3] capacity of another. Moreover, they show that this measurement reflects host-parasite interactions and facilitates predicting such interactions on a large scale.

Several experiments and field studies investigated microbial interactions (Krabberød *et al.*, 2017) like intraspecific and interspecific competition (Fredrickson & Stephanopoulos, 1981), predation (Guerrero *et al.*, 1986), parasitism (Chambouvet *et al.*, 2008), and mutualism (Gast & Caron, 1996). For example, Guerrero *et al.* (1986) observed and characterized two kinds of predatory bacteria and concluded that antagonistic relationships, such as primary consumption, predation, and scavenging, had already evolved in microbial ecosystems before the appearance of eukaryotes.

However, most marine microbial interactions are still unknown (Krabberød *et al.*, 2017). Like the ones mentioned above, most studies focused on relationships within a single or few species, which provides no insight into the complex system of ecological interactions occurring in microbial communities. Thus, when studying an ecosystem, the possible interactions between all microorganisms that constitute it should be included. Finally, to achieve a holistic understanding of the ecosystem, at least two types of data are needed:

- A *list of components* representing microorganisms;
- A *list of interactions* between these components in a spatiotemporal context.

[3] Biosynthesis is a multi-step process catalysed by specific proteins (enzymes), where substances (substrates) are converted into more complex ones (products) in living organisms.

Recent advances in omics technology have allowed to retrieve a large number of microorganisms, and to improve the knowledge of ocean microbiome ecological relevance. The technological advance allowed to identify and quantify the list of components (next section), which is used to predict a list of interactions via associations (Chapter 3).

Identifying and quantifying microorganisms

Many microorganisms remain unknown or poorly known because they are uncultured (Baldauf, 2008; Lewis *et al.*, 2020). This limits the access to microbial genomes, gene-expression patterns, and metabolism, i.e., the information needed to make inferences about microbial function in the ecosystem. This limitation is currently overcome by relatively new high-performance approaches, including genomics, transcriptomics, metagenomics, metatranscriptomics, and single-cell genomics.

The advances in omics tools resulted in a revolution in microbial molecular ecology (Dupont *et al.*, 2010; Hasin *et al.*, 2017). Briefly, the *genome* is the set of genetic material that every organism has, while the *gene-expression*[4] *patterns* describe which parts are being expressed and used within the cell. *Metabolism* comprises processes or reactions needed to maintain life, subsequently, it is the set of life-sustaining chemical transformations. DNA regions on the genome are copied into genomic information called RNA *transcripts*. This is referred to as *transcription*: DNA is transcribed into RNA. Specific RNA, so-called messenger RNA (mRNA), act as a blueprint to form proteins; a process called *translation*: mRNA is translated into proteins. Following these definitions, *genomics* is the study of genomes. *Metagenomics* investigates the collective set of genomes within a community. *Transcriptomics* concentrates on the transcriptome, the complete set of RNA transcripts present in a cell under defined conditions and their quantities. *Metatranscriptomics* examines the collective set of RNA transcripts within a community. Finally, if these studies are applied to a single isolated cell, we might refer to *single-cell genomics* and *single-cell transcriptomics*. Single-cell genomics and transcriptomics enable studying the extent and nature of genomic and transcriptomic heterogeneity (Macaulay & Voet, 2014). Technologies to obtain omics data are high-throughput and generate a massive amount of data. Thus, we need new algorithms to work with such data in an acceptable amount of computing time.

The advances in high-throughput technologies have made it possible to identify and quantify the list of components, that is, the microorganisms inhabiting the ocean (Li *et al.*, 2016). The procedure to obtain such a list of components includes four main steps.

[4] Gene: a region on the DNA (Deoxyribonucleic acid) that transcribes into RNA (Ribonucleic acid). This RNA has a direct function or gets translated into an amino acid sequence. The set of all RNA of a cell at a given moment is used to determine the expression profile.

Sampling: The sampling of marine microorganisms in the open ocean occurs from a ship. Water is collected and typically filtered for different size fractions. The sampling might take place across different sites or time points of the same site. The former approach leads to spatial data, the latter to temporal data. The time component allows to better infer interactions between species and model the microbial ecosystem (Li *et al.*, 2016). Li *et al.* state that, on the contrary of spatial data, temporal datasets provide a dynamic view of interactions, e.g., they can be used to infer the direction of dependency and therefore distinguish commensalism and competition (Li *et al.*, 2016). Together with the sampling, a large set of environmental variables are measured, such as temperature, salinity, and concentration of chlorophyll as a proxy for primary productivity (photosynthetic capacity), and many more.

Sequencing: Once the microorganisms from the desired size fraction have been filtered, we can extract their genetic material for sequencing. Sequencing is the process of determining the genomic information in which the components of the DNA, called bases, are extracted as a linear sequence. Metagenomics aims to sequence the genomic content from a community of microorganisms to obtain a fair representation of what is in a natural sample. It is usually carried out via *shotgun DNA sequencing*, a method to sequence long DNA fragments. The DNA is randomly cut into numerous small pieces that are filtered for their length and sequenced. Not necessarily the entire fragments are sequenced, but a subsequence on one or both ends, which can overlap if fragments are sufficiently small. The latter approach is called *paired-end sequencing*. The resulting subsequences are called *reads* and *paired-end reads (forward and reverse reads)*, respectively. They can be assembled, i.e., overlapped, to derive the DNA sequence from where they originated, through a bioinformatic process called *assembly*. The assembled reads compose a contiguous subsequences of the DNA (contig). The additional information obtained through paired-end sequencing, which is the distance between two paired-end reads, reveals contigs' orientation, and allows to assemble them into scaffolds (Figure 1).

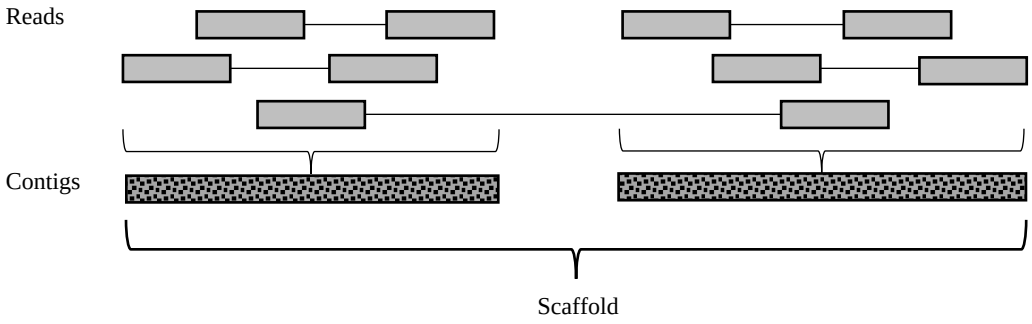


Figure 1: **Illustration of reads, contigs, and a scaffold.** Reads assemble to contigs. Contigs can constitute scaffolds with the help of paired-end reads.

Sequencing the entire genomic content of each microorganism in a community is costly and generally unfeasible. Thus, another method, the commonly used polymerase chain reaction (PCR) technique copies a specific DNA fragment whose ends are defined (Erlich, 1989). Some specific genes or regions of the DNA have been profoundly studied and give enough information to identify microbial species. Thus, a shorter region is often sufficient and referred to *marker* or *target genes*. In taxonomy studies, the favored (gold-standard) marker gene codes[5] for the small ribosomal subunit (Baker *et al.*, 2003; McNichol *et al.*, 2020), i.e., the *16S ribosomal RNA* gene for prokaryotes (archaea and bacteria), and the *18S ribosomal RNA* gene for eukaryotes, shortly referred to as 16S rRNA and 18S rRNA gene. These sequence data are widely used in molecular analysis to reconstruct the evolutionary history of organisms, because such regions have a low evolutionary rate, i.e., the sequence changes slowly over time. Moreover, the sequence is homologous between eukaryotes and prokaryotes since they share the same ancestor, but some sub-regions are variable, allowing to differentiate between different microorganisms that are closely related. This desired short DNA region coding for the 16S/18S rRNA gene is copied and multiplied, i.e., repeatedly synthesized during sequencing. We refer to it as the 16S/18S rRNA gene being *amplified* to generate thousands and millions of copies. Such a copy is called *amplicon* and typically consists of 200-450 bases.

Clustering: The concept of a biological species for microorganisms is not clear. One of the main difficulties related to the analyses of marker genes is to distinguish sequencing errors from true heterogeneity (Bharti & Grimm, 2021). One pragmatic solution resolves sequencing errors by clustering reads on the basis of a predefined identity threshold into groups called *Operational Taxonomic Units* (OTUs) (Westcott & Schloss, 2015). OTUs may or may not agree with biological species. However, OTUs provide a "unit" used for ecological analyses (Logares *et al.*, 2012). The number of OTUs depends on the similarity threshold (e.g., 97% vs. 99%), and on the clustering algorithm employed (Edgar, 2013; Sinclair *et al.*, 2015).

Another solution is a clustering-free approach, which uses a denoising approach on biological sequences before the introduction of amplification and sequencing errors (Tikhonov *et al.*, 2015). The sequencing errors are sufficiently controlled, so sequences can be exactly resolved, leading to a fine resolution and allowing to separate sequences into so-called *Amplicon Sequence Variants* (ASVs) (Callahan *et al.*, 2017). Using ASVs instead of "the threshold"-OTUs has its advantages, such as computational costs linearly scaling with size of the study, and the possibility of simple merging or comparison between independently processed data sets (Callahan *et al.*,

[5] Coding: a specific combination of three DNA components (bases, more precise nucleotides) translate into one component of a protein (called amino acid). This mapping is not a bijective function: there are four nucleotides, i.e., 64 combinations of triplets, but only 20 amino acids. Some combinations do not map to an amino acid and act as a stop of the translation. The translation scheme is called the genetic code. If a region of the DNA codes for a protein it means that the protein was built upon an mRNA copy from that DNA region.

2017). Callahan *et al.* (2017) state that replacing OTU with ASV makes marker-gene sequencing more precise, reusable, reproducible, and comprehensive.

Sequence abundance/count table: Once the OTUs/ASVs are defined, the amplicons are counted and represent the sequence abundance of the OTUs/ASVs in a sample. The collected sequence abundance data can then be further tabulated into a matrix. Numerous efforts have been made for sampling and they resulted in sequence abundance tables for marine microbial studies. Several expeditions provide samples from various locations worldwide or regionally, e.g., TARA Ocean (Karsenti *et al.*, 2011), Malaspina (Duarte, 2015), and Hotmix (open Mediterranean Sea and the adjacent Northeast Atlantic Ocean) (Martínez-Pérez *et al.*, 2017). In addition, several efforts have been made to obtain temporal data from microbial time-series, e.g., the San Pedro Ocean Time Series (SPOT) from the San Pedro Channel, off the coast of Los Angeles in southern California (Cram, Chow, *et al.*, 2015), the Banyuls Bay Microbial Observatory (SOLA) (Lambert *et al.*, 2021), and the Blanes Bay Microbial Observatory (BBMO) (Gasol *et al.*, 2016). SOLA and BBMO are both located in the North-Western Mediterranean Sea.

Here, we use in-house temporal and spatial data. Precisely, we had access to the ten-year BBMO time-series (Gasol *et al.*, 2016) and data from the two expeditions Malaspina-2010 (Duarte, 2015) and Hotmix (Martínez-Pérez *et al.*, 2017). BBMO sampled one coastal location in the North-Western Mediterranean Sea (Gasol *et al.*, 2016); Hotmix the open Mediterranean Sea and the adjacent Northeast Atlantic Ocean (Martínez-Pérez *et al.*, 2017); and the Malaspina expedition covered several oceans. The Malaspina expedition aimed to explore the global biogeography, diversity, functioning, and genetic interaction of deep-ocean and surface marine microorganisms (from small eukaryotes to prokaryotes and viruses), for example through 16/18S rDNA sequencing, coupled to meta-genomics and meta-transcriptomics. The dataset contains picoplankton distributed worldwide, including ocean surface (3 m depth), deep ocean (~4000m depth), and data obtained along the water column (vertical profiles at 11 stations with up to 7 depths from the surface to the deep ocean).

Analyzing and converting microbiome data into meaningful biological insights is challenging, but network-based approaches have the potential to help disentangle microbial interactions (Layeghifard *et al.*, 2017). The next chapter introduces and defines the mathematical aspects of networks. Before, we shortly introduce one specific microbial group.

Example group: Cyanobacteria

Microbial associations provide interaction hypotheses, which could lead to further investigations. This thesis focuses on general patterns but uses *Cyanobacteria*, the so-called blue-green algae, as an example group in Chapter 6). *Cyanobacteria* were selected because of their important ecological role as primary producers and their interactions with other organisms, e.g., to which they supply nitrogen (Scuito & Moro, 2015). Occurring in almost every habitat on Earth including

extreme environments (desert soils, glaciers, and hot springs) facing biotic and abiotic stresses, cyanobacteria produce a large array of metabolites including toxins (Scuito & Moro, 2015). Moreover, they require simple ingredients to grow and have a high growth rate. These characteristics allow a wide range of applications in nutrition, bioenergy, biotechnology, natural products, medicine, agriculture, and environment (Scuito & Moro, 2015; Zahra *et al.*, 2020). For example, the production of anti-cancerous cyanobacterial bioactive compounds by *Cyanobacteria* makes them useful in the pharma and healthcare sectors (Qamar *et al.*, 2021). Thus, cyanobacteria can have both positive and negative impacts on the environment and on human activities (Scuito & Moro, 2015). However, they have been far more critical to life on Earth.

Cyanobacteria are the first known oxygenic photosynthetic microorganisms. They contributed to the first rise in oxygen in the atmosphere of the Earth and shallow ocean, the so-called Great Oxidation Event. However, it is unknown when they first appeared and past evidence based on molecular fossils was demonstrated invalid (Rasmussen *et al.*, 2008). Thus, the most substantial indirect evidence for the appearance of oxygenic cyanobacteria is the rapid accumulation of atmospheric oxygen between 2.45 and 2.32 x 10^9 years ago (Bekker *et al.*, 2004). This event is one of the most significant changes in Earth history, setting the stage for extensive transformations in ocean chemistry and the evolution of multicellular life (Pufahl & Hiatt, 2012). Pufahl and Hiatt (2012) concluded that it is the "*utmost expression of co-evolution between the geosphere and biosphere*". The geosphere provided the chemical building blocks and ecological niches for early life. The biosphere provided oxygen, which changed the nature of weathering; nutrient cycling; mobility of redox-sensitive elements (like iron and uranium); and environmental stresses that pushed life along new evolutionary pathways (Pufahl & Hiatt, 2012).

Eukaryotic photosynthesis originated from the endosymbiosis of cyanobacterial-like organisms. In 1967, Lynn Margulis (then known as Sagan) revived the forgotten theory of organelle endosymbiont origin (Sagan, 1967), proposing the endosymbiotic origin of chloroplasts from *Cyanobacteria* (Mereschkowsky, 1905, 1910; Martin & Kowallik, 1999; Kowallik & Martin, 2021), and mitochondria from *Alphaproteobacteria* (Wallin, 1927). A decade later, Margulis' hypothesis was tested with phylogenetic trees from protein sequences (Schwartz & Dayhoff, 1978). Conserved genes in chloroplasts and mitochondrial genomes clustered with cyanobacterial and alphaproteobacterial genes, respectively. However, questions like when, where, and how eukaryotic cells evolved remain unclear (López-García *et al.*, 2017).

Cyanobacteria have various shapes and sizes: from picoplankton (~500 nanometers in diameter, invisible in a conventional light microscope), to relatively large cells forming chains that are visible to the naked eye (Falkowski, 2015). They have symbiotic interactions with eukaryotes, e.g., they are prey to them. A literature-based eukaryote interaction database[6], called

[6] Database downloaded 15th October 2019 and no updated version was available by 1st of June 2021.

PIDA (Bjorbækmo *et al.*, 2019), lists 77 interactions involving Cyanobacteria. All of them belonged to *Cyanophyceae* and included the genera *Synechococcus* (16 entries) and *Prochlorococcus* (only one entry). Although these two genera comprise only 22% of the 77 database entries, they were among the most abundant microorganisms in our datasets. We also found the genus *Cyanobium* within our data, but no entry for it in the interaction database. Given the importance of *Cyanobacteria* and the discrepancy between their known and potential (network-inferred) associations, they are an ideal candidate to highlight and investigate further.

Final remarks

- ⇒ Marine microorganisms are crucial for the functioning of the ecosystem.
- ⇒ To understand the ecosystem, we need to i) identify and quantify the microorganisms; and ii) determine their interactions.
- ⇒ Advances in omics high-throughput technologies supported microbial investigations by identifying and quantifying microorganisms.

Chapter 2 Graph-theoretic aspects

The interplay between microorganisms can be translated into mathematical language using networks. Network theory is part of graph theory. A *network* is a graph. In this thesis, I mainly used *simple undirected graphs*:

- undirected means that edges are without a direction, no start nor end node;
- there are no loops, meaning that no edge can connect a node to itself;
- the graph does not contain multiple edges, meaning that there is either one or no edge between two nodes.

The graph elements (nodes, edges) can have attributes assigned to them. For example, a node attribute may be the microbial taxonomic group, and an edge attribute the strength of the microbial association the edge represents. Thus, graphs are an ideal tool to comprise concisely much information within one object, the graph. Moreover, graph-based characteristics for nodes, edges, and emergent properties of the graph as a whole can be determined.

Basic definitions

The graph

A ***graph*** G is a pair (V, E), with V representing the set of nodes (an element of V is called a ***node***) and E representing the set of edges (an element of E is called an ***edge***). An ***undirected*** edge $e = (v_1, v_2)$ has the nodes v_1 and v_2 at its ends. A ***directed*** edge $e = [v_1, v_2]$ has a start node v_1 and an end node v_2. Here, V is a finite set. It constitutes the set of E using pairs from V, i.e., E is a subset of (V, V).

Definition 1

A ***graph*** $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ is defined through a set of nodes, $\mathrm{V} = \{\mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{n}\}$, and a set of edges between nodes, $\mathrm{E} = \{\mathrm{e}: \mathrm{e} = (\mathrm{v}_\mathrm{i}, \mathrm{v}_\mathrm{j}) = (\mathrm{v}_\mathrm{j}, v_i)$ with $\mathrm{v}_\mathrm{i}, \mathrm{v}_\mathrm{j} \in \mathrm{V}$, and $\mathrm{i} \neq \mathrm{j}\}$.

Two nodes are ***adjacent*** if they are connected through an edge. An edge and a node on that edge are ***incident***. A graph can be represented through an ***adjacency matrix*** in which the columns and rows represent the nodes. If two nodes connect through an edge, the entry in the matrix is 1; otherwise, it is 0. If the graph is undirected, i.e., all edges are undirected, the adjacency matrix is symmetric. Further, a graph can be represented through an ***incidence matrix*** in which the columns represent the edges and rows represent the nodes. If an edge and a node are incident, the entry in the matrix is 1; otherwise, it is 0. Lastly, a graph can also be represented through an

edge list. If the matrix is *sparse*, i.e., the adjacency/incidence matrix contain many zeros, the network is usually stored as a list to reduce memory requirement.

Neighbors of nodes

Given a node v_i is adjacent to node v_j. Then, node v_j is called the ***neighbor*** of node v_i. All neighbors of v_i constitute the ***neighborhood*** of node v_i, the set $N_v = \{v_1, \dots, v_k\} \subseteq V$. The number of neighbors is the ***degree*** of a node in an undirected network. In a directed network, the number of edges pointing to the node is the in-degree, and the number of edges pointing away the out-degree. Let n_k be the number of nodes in the network with degree k. Then the ***degree distribution*** p_k is the probability that a randomly chosen node in graph G has degree k.

Definition 2

The ***degree*** of a node v in V of graph G is the number of edges attached to it, i.e., the cardinality of its neighborhood. Let $|V| = n$ and n_k is the number of nodes with degree k. Then, the ***degree distribution*** $P(k)$ of the network is the fraction of nodes with degree k:

$$P(k) = \frac{n_k}{n}.$$

Paths and shortest paths

A ***path*** between two nodes is a sequence of edges that connect them, i.e., the edges needed to traverse the graph to get from one node to another node. The number of edges is the ***length*** l of the path. A path of minimum length, l_{min}, is called a ***shortest path***. The shortest path between two nodes may not be unique. The length of the shortest path between two nodes is the ***distance*** of these two nodes: $d(v_i, v_j) = l_{min}$. We say a network is ***connected*** if there exists a path between all pairs of nodes, i.e., it exists a path $(v_i = v_1, v_2, \dots, v_l = v_j)$ with length $l \geq 0, l \in \mathbb{N}$ for all disjunct pairs of nodes $\{v_i, v_j\}$ in V with $i \neq j$.

Connected and unconnected graphs

A graph can be connected or unconnected. A ***connected graph*** contains a path between any two nodes. An ***unconnected graph*** contains at least one node that is not connected to at least one other node. A ***complete graph*** is a special case of a connected graph: each node connects to every other node through one edge, i.e., all shortest paths have length 1.

Subgraph and induced subgraph

Given graph G with node set V and edge set E, graph G' is called a ***subgraph*** of graph G if its node-set V' is a subset of V and its edge-set E' a subset of E. All nodes incident to the edges in E'

also appear in V'. In contrast, not all edges incident to the nodes in V' must appear E'. On the contrary, an ***induced subgraph*** G'' of graph G has a subset V'' of the node-set V, and all incident edges appearing in graph G also appear in the subgraph G''.

Clique

A ***clique*** is a fully connected (complete) subgraph, i.e., each node connects to all other nodes through one edge.

Triplet

A ***triplet*** is a tiny graph consisting of three nodes connected via two edges (***open triplet***) or three edges (***closed triplet***). A closed triplet is a clique.

Global graph metrics

There are redundancies between global graph metrics grouping them into four so-called redundancy clusters (Jamakovic & Uhlig, 2008). Thus, for graph characterization, we selected one from each redundancy cluster: average path length (distance cluster), transitivity (degree cluster), edge density (intra-connectedness), and assortativity based on node degree (inter-connectedness). The latter can also be determined on the basis of a nominal classification, e.g., using the two taxonomic groups eukaryote and prokaryote. The global graph metrics are defined below.

Average Path Length

Definition 3

The ***average path length*** is the average length of all possible shortest paths in the graph. Therefore, it is a positive number.

Transitivity

The ***transitivity***, or ***clustering coefficient***, measures the probability that two neighbors of a node are also connected. Here, we define transitivity through the ***global clustering coefficient*** C of a graph. It is defined as the number of closed triplets over the total number of triplets, i.e., open and closed triplets:

$$C = \frac{number\ of\ closed\ triplets}{number\ of\ triplets}.$$

Further, let N_i be the neighborhood of node v_i and $k_i = |N_i|$ be the number of its neighbors. The ***local clustering coefficient*** C_i of node v_i is the proportion of edges between the nodes within its neighbors divided by the potential number of edges between them, k_{max}. Then the local clustering coefficient of node v_i is computed as

$$C_i = \frac{|\{(v_h, v_j): v_h, v_j \in N_i \ \wedge (v_h, v_j) \in E\}|}{k_{max}}.$$

For an undirected network $k_{max} = \frac{1}{2}k_i(k_i - 1)$, and directed network $k_{max} = k_i(k_i - 1)$. Alternatively, to the global clustering coefficient, the overall level of clustering in a network can also be measured through the ***average clustering coefficient***, which is the average of the local clustering coefficients of all nodes (Watts & Strogatz, 1998):

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i.$$

Definition 4

The ***transitivity***, measuring how well nodes cluster together, is the ratio of closed triplets to all triplets, i.e., closed and open triplets. The transitivity ranges between 0 and 1.

Edge density

Definition 5

The ***edge density***, measuring how well the graph is connected, is the ratio of observed to possible edges. Given a graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, the edge density is $\frac{2|E|}{|V|\cdot(|V|-1)}$ with $|\mathrm{V}|$ and $|\mathrm{E}|$ representing the cardinality of the node and the edge set, respectively. The edge density ranges between 0 and 1.

Assortativity

The fourth metric, the assortativity coefficient, measures the relationship between nodes (Newman, 2002). The following explanations are adapted from (Newman, 2002). In general, assortativity ranges between -1 and 1. It is positive if similar nodes (on the basis of an external property) tend to connect, and negative otherwise. If it is 1, the graph has perfect assortative mixing patterns (assortative graph), zero indicates non-assortativity, and -1 complete disassortativity. The popular external metric is the node degree.

Defining assortativity is not straightforward. Above, we defined the degree distribution (Definition 2). However, if we randomly choose an edge in E of graph G and consider the incident

node v, its node degree is not distributed according to the above-defined degree distribution. It is biased in favor of high-degree nodes because more edges end at a high-degree node than a low-degree node. Subsequently, the degree distribution for the node at the end of a randomly chosen edge is proportional to kp_k. If we use one edge to get to a node, then the other "remaining" edges comprise the *remaining degree*.

Definition 6

The ***remaining degree*** of a node v in V of graph G is the number of edges attached to it minus 1. It is distributed in proportion to $(k+1)p_{k+1}$. Thus, the normalized distribution q_k of remaining degrees is $q_k = \frac{(k+1)p_{k+1}}{\sum_j jp_j}$.

An edge connects two nodes. Then the joint probability distribution e_{jk} of the remaining degrees of two nodes incident to a randomly chosen edge, obeys the sum rules $\sum_{j,k} e_{jk} = 1$, and $\sum_j e_{jk} = q_k$. In a disassortative graph, e_{jk} takes the value $q_j q_k$. In an assortative network, e_{jk} differs from that value. The amount of assortative mixing in a graph can be quantified by averaging the differences: $\sum_{j,k} jk(e_{jk} - q_j q_k)$. This is the so-called connected degree-degree correlation function. It is zero for no assortative mixing. It is positive for assortative and negative for disassortative mixing. The value is normalized by dividing it by its maximal value, obtained on a perfectly assortative graph. This value is equal to the variance $\sigma_q^2 = \sum_k k^2 q_k - [\sum_k k q_k]^2$ of the distribution q_k.

Now, we can define assortativity on the basis of the external node characteristic "node degree". In addition, we also define assortativity with a nominal characteristic.

Definition 7

The ***assortativity*** **(degree)** is the Pearson correlation coefficient of degrees at either ends of an edge: $r = \frac{1}{\sigma_q^2} \sum_{j,k} jk(e_{jk} - q_j q_k)$.

The ***assortativity*** **(nominal)** is the assortativity for categorical labels of nodes:

$$r = \frac{\sum_i e(i,i) - \sum_i a(i)b(i)}{1 - \sum_i a(i)b(i)},$$

where $a(i) = \sum_j e(i,j)$ and $b(j) = \sum_i e(i,j)$, $e(i,j)$ is the fraction of edges connecting nodes of type i and j, and, subsequently, e(i,i) is the fraction of edges connecting nodes of the same type i.

In our work, we use both assortativity on the basis of degree and a nominal classification. Precisely, we use the taxonomic classification into eukaryote (Euk) and prokaryote (Prok).

Moreover, our graph G is a simple undirected graph leading to $e(i,j) = e(j,i)$. Thus, the formula for assortativity (nominal) simplifies to

$$r = \frac{e(Euk,Euk)+e(Prok,Prok)-2e(Euk,Prok)^2}{1-2e(Euk,Prok)^2}.$$

Local graph metrics

Global and local graph metrics characterize graphs. Above we described global graph metrics (edge density, average path length, transitivity, and assortativity). However, global graph metrics disregard local structures' complexity, and topological analyses should include local graph metrics (Espejo *et al.*, 2020). Local-topological metrics may use either motifs or graphlets. Motifs are subgraphs, and graphlets are induced subgraphs. Thus, the latter preserves all connections among nodes contrary to motifs.

Graphlets are small connected induced subgraphs of a graph (Pržulj *et al.*, 2004), i.e., a graphlet considers all edges for a given set of nodes. The smallest graphlet contains two nodes and one edge (G0). Graphlets with three nodes have two edges (G1) or three edges (G2). G1 and G2 are an open and a closed triplet, respectively. While the roles of a node in G2 are comparable (each node has two neighbors), there are two roles within G1: being connected to two nodes or being connected to one node.

Mathematically, an open triplet has two ***automorphism orbits***: orbit 1, the node is connected to one other node (black node in G_1 in Figure 2), and orbit 2, the node is connected to two other nodes (white node in G_1 in Figure 2). Let v be a node in a graph G. Then, the automorphism orbit of v is the set of nodes of G that can be mapped to v by an automorphism (an isomorphism of a network with itself) (Yaveroğlu *et al.*, 2014). Nodes of the same automorphism orbit within a graphlet are indicated in the same color in Figure 2, e.g., the two black nodes in G1. Any bijection of nodes belonging to the same automorphism orbit preserves node adjacency. Orbits define the relative position of nodes with respect to other nodes in the graphlet. There are 15 orbits among the nine 2- to 4-node graphlets (Figure 2).

Orbits extend the node degree through the so-called ***graphlet degree vector***. Let C_i be the i-th graphlet degree of a node with i indicating the orbit (see orbit 0 to 14 in Figure 2). For example, C_0 is the degree of a node as it counts the number of times the node is *touched* by orbit 0. Similarly, C_2 is the graphlet degree for orbit 2 and C_3 the graphlet degree for orbit 3. Graphlets provide a complete description of local graph topology (Espejo *et al.*, 2020). Although considering all graphlet sizes may complete the description of graph topology, it would be computationally expensive.

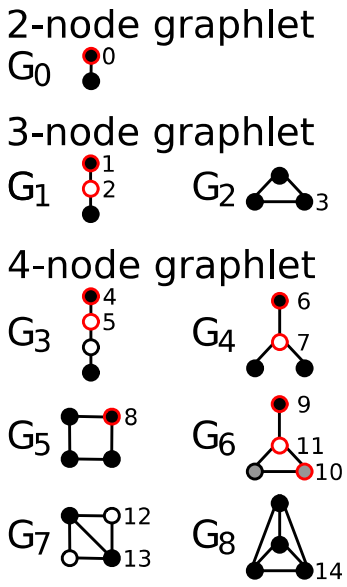


Figure 2: **The nine 2- to 4-node graphlets G_0, …, G_8.** Nodes of the same color belong to the same role (automorphism orbit). Some orbits are redundant as their counts in a network can be derived from the counts of other orbits. The 11 red circles indicate non-redundant orbits. The selection of non-redundant orbits is not unique. I adapted this figure from Figure 1d in (Yaveroğlu *et al.*, 2014).

Some orbits can be computed from others, leaving 11 non-redundant orbits among the nine 2- to 4-node graphlets (Yaveroğlu *et al.*, 2014). For example, consider a node with C_0 neighbors (orbit 0 in graphlet G_0). The neighbors of the node can either be connected (a closed triplet, then the node touches orbit 3 in graphlet G_2) or they are not connected (an open triplet, then the node touches orbit 2 in graphlet G_1). With C_0 neighbors, there are $\binom{C_0}{2} = \frac{C_0(C_0-1)}{2}$ pairs of neighbors that are either connected or not connected. Then, $\binom{C_0}{2} = C_2 + C_3$. Thus, one of the three orbit degrees (C_0, C_2, or C_3) is redundant as it can be computed from the other two. Yaveroğlu *et al.* (2014) indicate this simple and another 16 independent orbit redundancy equations (i.e., they cannot be derived from other equations).

The selection of non-redundant orbits is not unique as there are several ways to choose non-redundant orbits (Yaveroğlu *et al.*, 2014). One set of 11 non-redundant orbits is indicated in red in Figure 2. Yaveroğlu *et al.* (2014) proposed to analyze these 11 non-redundant orbits to characterize graph structure and to determining graph (dis)similarity. Shortly, given there are n nodes in graph G. First, for each node, the graphlet degree vector considering the 11 non-redundant orbits is determined. The vectors are comprised to a $n \times 11$ matrix, i.e., it has n rows and 11 columns. The number of rows is equal to the number of nodes in the graph. The number of columns is equal to the number of the considered graphlet degrees.

The $n \times 11$ matrix listing all graphlet degree vectors is used to compute a new statistic for the graph by computing the Spearman's correlation coefficients between all pairs of columns, which aims to exploit the existence of monotonic correlations between the non-redundant orbits. If one of the non-redundant orbits does not appear in the graph, the corresponding column contains only 0. The Spearman's correlation coefficient cannot be computed for these orbits but

Yaveroğlu *et al.* (2014) overcame this computation problem by adding a pseudo row to the matrix, i.e., a dummy graphlet degree vector, $[1, 1, \cdots, 1]$ resulting in a $(n + 1) \times 11$ matrix. Orbits for which all graphlet degrees are 0 will correlate perfectly resulting in a Spearman correlation coefficient of 1. Further, these orbits will not correlate with non-zero orbits resulting in a Spearman correlation coefficient of 0.

This results in a symmetric 11×11 matrix, which is called ***graphlet correlation matrix (GCM)*** (Yaveroğlu *et al.*, 2014). Thus, regardless of the number of nodes in a network, the topology of a graph can be summarized into an 11×11 matrix with values in the interval $[-1, 1]$ (Yaveroğlu *et al.*, 2014). That is, the GCM encodes the topology of a network using correlations between node properties contained in their non-redundant orbit counts (Yaveroğlu *et al.*, 2014).

Generally, different real and model graphs have very different orbit dependencies resulting in different GCMs according to Yaveroğlu *et al.* (2014), who used the GCM to define a distance metric to measure graph dissimilarity. We indicate the matrix element in row i and column j with $GCM(i, j)$. Further, two graphs G_1 and G_2 are given via their graphlet correlation matrices GCM_{G_1} and GCM_{G_2}. Then, we can measure graph dissimilarity via the so-called ***graphlet correlation distance (GCD)***, which is computed as the Euclidean distance of the upper triangle values of GCM_{G_1} and GCM_{G_2}:

$$GCD(G_1, G_2) = \sqrt{\sum_{i=1}^{10} \sum_{j=i+1}^{11} \left(GCM_{G_1}(i, j) - GCM_{G_2}(i, j) \right)^2}.$$

Thus, GCD encodes information about local network topology and provides a non-negative distance between two graphs, with 0 indicating identical graphs and the greater the distance the less similar (more dissimilar) are two graphs (Yaveroğlu *et al.*, 2014).

Different types of graphs

There are different types of graphs; the four main types of graphs, according to (Layeghifard *et al.*, 2017) are i) regular graph, ii) random graph, iii) small-world graph, and iv) scale-free graph. Within a ***regular graph***, each node has the same number of edges. In a ***random graph***, nodes randomly connect. Within a ***small-world graph***, most nodes can be reached from any other node through a path of short length. Let the average shortest path length be approximately the same as one of a random graph. Then the random and small-world graphs differ in their average clustering coefficient as the one of a small-world graph is higher than the one of a random graph. ***Scale-free graphs*** are characterized through their degree distribution of nodes, which follow a power law, i.e., for a constant $\gamma > 0$ is $P(k) \sim k^{-\gamma}$ (Barabási & Albert, 1999).

Moreover, let the node-set V be divided into two disjunct sets, and edges only connect nodes from one set with nodes from the other disjunct set, i.e., there are no edges connecting nodes within the same set. Then the graph is called ***bipartite***. Bipartite graphs are often used to connect two types of nodes, e.g., hosts with parasites.

A ***minimal spanning tree*** is a special case of a connected subgraph of a graph. The edges of the graph are assigned an edge weight. The edges of the minimal spanning tree are chosen so that i) all nodes are connected through a path, and ii) the sum of edge weight is minimal.

In order to define a temporal graph, we first define a more generalized object, a ***multilayer graph***, of which temporal graphs are a special case. A multilayer graph comprises several layers. Each layer is a simple graph with its nodes and edges. Moreover, the graphs in different layers can be connected, allowing edges between layers similar to edges in bipartite graphs. On the basis of previous work (Bianconi, 2018), we define a general multilayer graph as follows:

Definition 9

A ***multilayer graph*** M is given through the triple $\mathcal{M} = (Y, \vec{G}, \mathcal{G})$.

- Y is the set of layers $Y = \{\alpha | \alpha \in \{1,2,\cdots,M\}\}$, with M indicating the total number of layers.
- $\vec{G}$ indicates the ordered list of graphs characterizing the interactions within each layer $\alpha = 1,2,\cdots,M$, i.e., $\vec{G} = (G_1, G_2, \cdots, G_\alpha, \cdots, G_M)$, where $G_\alpha = (V_\alpha, E_\alpha)$ is the graph in layer α. The set of nodes of layer α is V_α and the set of edges E_α. These edges are also called ***intra-edges***.
- $\mathcal{G}$ is a $M \times M$ list describing the edges between layers. Elements of $\mathcal{G}$ are denoted as $\mathcal{G}_{\alpha,\beta}$, which are given by $\mathcal{G}_{\alpha,\beta} = (V_\alpha, V_\beta, E_{\alpha,\beta})$ for each $\alpha < \beta$ and $\alpha, \beta \in \{1,2,\cdots,M\}$. Here, $\mathcal{G}_{\alpha,\beta}$ indicates the bipartite graph with nodes sets V_α and V_β, and the edge set $E_{\alpha,\beta}$. These edges are also called ***inter-edges*** and connect the nodes between the layers α and β.

We can employ the framework of multilayer graphs to define temporal graphs. ***Temporal graphs*** are a special case of multilayer graphs: the layers represent discrete time points, and the sets of nodes in each layer have the same type, e.g., all nodes in each layer are microorganisms.

Definition 10

A ***universe*** $\mathcal{U}$ is a collection of all considered entities.

Definition 11

A ***temporal graph***, $T = (Y, \vec{G}, \mathcal{G})$, is a special case of a multilayer graph in which the layers $Y = \{\alpha | \alpha \in \{1,2,\cdots,M\}\}$ correspond to discretized time points, and the set of nodes V_α with $\alpha \in \{1,2,\cdots,M\}$ are subsets of the same universe $\mathcal{U}$.

The third variable $\mathcal{G}$, representing the list of bipartite graphs in a temporal network, allows to include time-delayed associations. Determining time-delayed edges may not be possible in each application. If time delay is not considered, a simpler version of the temporal graph can be given through $\mathcal{M} = (Y, \vec{G})$. Alternatively, temporal networks can be defined by extending the node and edge set by the temporal dimension, i.e., assigning a property indicating presence and absence for each time point.

Final remarks

- ⇒ Graphs used to model a specific system are often referred to as networks.
- ⇒ Ecosystems, such as marine microbial ecosystems, are complex systems that can be modeled through networks.
- ⇒ The toolbox of graph-theoretic concepts is diverse and will benefit marine microbial interaction studies.
- ⇒ In this thesis, we mainly consider simple undirected networks.
- ⇒ In subproject 2, we use simple undirected subnetworks (not induced) to determine a temporal network.
- ⇒ Networks can be characterized through global and local network metrics.
- ⇒ The degree-vector can be extended by the graphlet degree vector.
- ⇒ Graphlet degree vectors are the basis for network comparison based on local patterns.
- ⇒ In subproject 3, we use graphlet based network similarity to compare subnetworks, and a minimal spanning tree as a comprehensive representation of network similarities.

Chapter 3 Microbial association networks

Microorganisms are involved in various ecological interactions (Table 1), which can be represented as an ecological network. Most interactions are unknown (Bjorbækmo *et al.*, 2019). Given n microorganisms, there are $\binom{n}{2} = \frac{n(n-1)}{2}$ pairs (potential interactions). For instance, if $n = 2000$, there would be 199000 pairs, which are too many to prove each one of them experimentally in a lab (given the microorganisms would be culturable). Omics-technologies allow to identify and quantify microorganisms generating profiles of sequence abundances for each sample, and the table of microbial sequence abundances. Such table allows to determine the microbial associations constituting a network. In contrast to ecological networks, association networks are limited by interpretational challenges since they cannot be directly observed (Röttjers & Faust, 2019), and tools used to infer associations display high error rates (Weiss *et al.*, 2016). Thus, microbial association networks provide (ecological) interaction hypotheses, which need further validation to obtain ecological networks.

Although the inference, analysis, and interpretation of marine microbial association networks encounter several challenges (see Chapter 4 and a very recent perspective (Faust, 2021)) and require to draw conclusions with care (Faust & Raes, 2012), networks are a versatile tool in microbial investigations (Faust & Raes, 2012):

- Network inference is generic, i.e., the same method can be applied to different data such as species or genes;
- Different types of data can be integrated, e.g., including microorganisms and environmental factors in one network;
- Different properties can be identified, e.g., relating to single nodes or associations, group of nodes, and the network as a whole.

Moreover, network analyses help to disentangle the structure of complex microbial communities across temporal and spatial scales (Barberán *et al.*, 2012; Fuhrman *et al.*, 2015).

In microbial ecology, most association networks comprise pairwise correlations or other mathematical relationships (Fuhrman *et al.*, 2015). Network-based analysis may detect essential microorganisms, e.g., by determining highly connected nodes (***hubs***), or groups of highly connected nodes (***modules***), which may represent groups of microorganisms of the same niche or biogeochemical process. In addition, networks display emergent properties related to characteristics of the community (Fuhrman *et al.*, 2015). Similar network topology for different networks may reveal common organization principles for different complex systems (Zhou *et al.*, 2010).

For instance, random networks follow a Poisson distribution (Erdös & Rényi, 1960). Yet, biological networks are clearly not random (Chaffron *et al.*, 2010; Steele *et al.*, 2011). For most

biological networks, the node degree distributions follow a power law (Barabási & Albert, 1999), at least in part (Khanin & Wit, 2006). Khanin and Wit (2006) indicate that most biological networks are not totally scale-free. Instead, they might better be described as following a truncated power law (suggesting scale-free behavior but only over a part of the network), while certain scale-free features, such as small world and centrality properties, hold true (Khanin & Wit, 2006). There are frequently many nodes with a low degree and few nodes with a high degree (hubs) (Khanin & Wit, 2006).

Networks with a small average shortest path are referred to as small-world networks (Watts & Strogatz, 1998). In small-world networks, most of the nodes can be reached from every other node through a shortest path with small length, i.e., small number of steps. Since a small-world pattern allows rapid communication among different components within a system, the system can respond to or is quickly affected by perturbations, such as environmental changes (Zhou *et al.*, 2010). Yet, high modularity can minimize effects on the whole system by containing perturbations at a local level (Kitano, 2004). In turn, a modules' hierarchical organization ensures a quick communication between modules and network hubs (Zhou *et al.*, 2010).

Previous studies of marine microbial association networks

Microbial association networks are popular exploratory tools deriving hypotheses from massive datasets (Röttjers & Faust, 2018). A network is considered a representation of a system that aggregated over some time (Steele *et al.*, 2011; Chow *et al.*, 2013, 2014; Cram, Xia, *et al.*, 2015; Needham *et al.*, 2017; Parada & Fuhrman, 2017), or a set of spatial samples (Lima-Mendez *et al.*, 2015; Milici *et al.*, 2016). Previous work characterizes ecological links between marine bacteria (Chow *et al.*, 2013; Cram, Xia, *et al.*, 2015) and eukaryotes (Milici *et al.*, 2016), archaea (Steele *et al.*, 2011; Parada & Fuhrman, 2017), and viruses (Chow *et al.*, 2014; Needham *et al.*, 2017). Another study includes organisms from viruses to small metazoans (Lima-Mendez *et al.*, 2015). Moreover, datasets along the water column allow to investigate within- and between-depth relationships (Cram, Xia, *et al.*, 2015; Lima-Mendez *et al.*, 2015; Parada & Fuhrman, 2017).

Previous studies identify associations among ecologically essential taxa, such as potential synergistic or antagonistic relationships, and possible keystone species and niches (Steele *et al.*, 2011; Chow *et al.*, 2013). Moreover, studies find more associations between microorganisms than between microorganisms and environmental factors, which indicates the dominance of microbial relationships over associations between microorganisms and environmental factors (Steele *et al.*, 2011; Lima-Mendez *et al.*, 2015). Finally, a previous study identifies associations that were driven by the environment, and determined regional associations (Lima-Mendez *et al.*, 2015).

Studies employing networks contribute to our understanding of marine microbial interactions. Networks are a great tool to handle the vast number of microorganisms and their connections,

explore potential microbial interactions, and elucidate patterns of microbial ecosystems. However, network construction is not straightforward.

Network construction

There are several methods inferring associations (list of interactions requirement) that use microbial sequence abundance data (list of components requirement). They vary in efficiency, accuracy, speed, computational requirements, and span from simple pairwise Pearson or Spearman correlation measures to Gaussian graphical models (Layeghifard *et al.*, 2017). Faust and Raes (2012) state that, on the basis of correlations and anti-correlations of components, it is possible to build networks and generate hypotheses about which components (microorganisms) may positively or negatively interact. Due to the current limitations of the data and network construction tools, it is suggested to remove extremely rare ASVs before network construction. Moreover, Weiss *et al.* suggest to lower the usually corrected p-value from 0.05 to 0.001, for a higher precision (Weiss *et al.*, 2016).

No ideal (gold-standard) method for network construction exists, and some tools are better suited than others for the specificities of a dataset (e.g., temporal vs. spatial, and homogenous vs. heterogeneous). Benchmarking the performance of eight correlation techniques (Weiss *et al.*, 2016), some methods perform better than others but there is still a considerable need for improvement. The investigation reveals that different tools infer significantly different numbers and types of edges for the same data, and they generally detect dissimilar edges (Weiss *et al.*, 2016). For all datasets/models tested by Weiss *et al.* (2016), two tools demonstrate on average an edge overlap of 31.5%. The low overlap suggests that the techniques may have different strengths and weaknesses in response to the diverse challenges presented by microbiome data, as mentioned in more detail in Chapter 4.

Since various metrics and also different tools have different strengths and weaknesses, they may also detect different functional relationships. An ensemble method with ReBoot procedure to assess significance is implemented in CoNet, which allows the combination of several metrics including similarity, dissimilarity and correlation measurements (Faust & Raes, 2016). Moreover, many tools have been designed to reduce the compositionality bias of microbial sequence data, e.g., SparCC (Friedman & Alm, 2012) and SPIEC-EASI (Kurtz *et al.*, 2015). In contrast to correlation techniques, which have been compared in (Weiss *et al.*, 2016), other tools are based on probabilistic graphical models, e.g., SPIEC-EASI (Kurtz *et al.*, 2015) and FlashWeave (Tackmann *et al.*, 2019). Hirano and Takemoto (2019) compared six correlation-based and three graphical model-based methods. For instance, the correlation-based methods included Pearson's correlation, Spearman's correlation and SparCC (Friedman & Alm, 2012), and the graphical model-based methods included SPIEC-EASI (Kurtz *et al.*, 2015). The

comparison (Hirano & Takemoto, 2019) indicated that i) compositional-data methods may not be more efficient than the classical methods; ii) graphical model-based methods may not be more efficient than the correlation-based methods; iii) interaction patterns in dense networks are difficult to predict; iv) interaction patterns in heterogeneous networks are the most difficult to predict while those in small-world networks, which are homogenous, are the easiest; v) tool performance increased with more samples and plateau for over 200 samples; iv) interaction types affect tool performance as also shown in (Weiss *et al.*, 2016).

Various tools exist but no best tool. We tested several tools (e.g., SparCC, SPIEC-EASI, MICtools, and CoNet) and selected two tools in consideration of our datasets: Within the three main subprojects, we used the tool eLSA (Xia *et al.*, 2011, 2013) for the temporal BBMO data and the tool FlashWeave (Tackmann *et al.*, 2019) for the spatial data compilation from Malaspina and Hotmix samples.

To construct a network from the temporal data set, we used eLSA (Xia *et al.*, 2011, 2013). It aims to capture time-dependent associations (possible time-shifted) between microorganisms and between microorganisms and environmental factors. First, to obtain normally distributed data, the tool normalizes the data on sequence abundance using normal score transformation (Li, 2002; Ruan *et al.*, 2006). Second, it calculates the association strength via statistical correlation (Pearson correlation coefficient) between any pair of components, including the maximum shifts allowed (time-delay), using dynamic programming. The maximum score of all subsequences within some predefined time delay is the so-called local similarity score (LS) (Ruan *et al.*, 2006). More specifically, eLSA determines the best start and end of the association in time for both association partners. A time delay represents a directed association. However, if the sampling is considered not suitable to allow time delay, the resulting network is undirected. Third, the tool determines the statistical significance by p-values using a permutation test. It randomly shuffles the components of the original data and recalculates the LS score for the pairs. Then, it approximates the p-value by the fraction of permutation scores that are larger than the original score. In addition, multiple testing corrections can be applied (q-values). Significant associations constitute the (and our preliminary) network.

For our spatial heterogeneous dataset compilation, we used FlashWeave (Tackmann *et al.*, 2019). After normalizing the data (centered log-ratio, clr transformation), FlashWeave determines pairwise associations (via a neighborhood search for each target variable). Next FlashWeave searches for conditional dependence between nodes, i.e., it removes associations between conditionally independent variables. Moreover, for datasets containing numerous zeros, FlashWeave is a tool of choice since it can handle sparse data and ignores matching zeros when computing associations (Tackmann *et al.*, 2019).

Important nodes

Microorganisms exerting a high impact on the structure and functioning of the ecosystem are referred to as *keystone* species. Keystones that are observed across different environments and studies may help microbial ecologists to explain the unexplained variation in ecosystem processes (Banerjee *et al.*, 2018). There are only few experimentally confirmed microbial keystones because the classical experimental validation consists in studying the effects of the removal/addition of putative keystone and is difficult to perform (Röttjers & Faust, 2019).

Networks may be used to infer key microorganisms via important nodes. However, importance in the network does not automatically translate into importance in the microbial system. For instance, highly abundant microorganisms are usually regarded as important by microbial ecologists. Marine microbial association networks are (at least in part) scale-free, with many low-degree nodes and a few high-degree nodes (called *hubs*). Hubs are important for network architecture. However, node degree does not depend on the abundance of the microorganisms that the node represents and abundance is generally very low for the highest degree nodes (Lima-Mendez *et al.*, 2015; Röttjers & Faust, 2018; Krabberød *et al.*, 2021).

Besides the number of associations (degree), other metrics and their combination may be used to infer important nodes (Berry & Widder, 2014). For instance, betweenness centrality reflects how important a node is for the connectivity of a network, by measuring how often it appears in the shortest path of two other nodes. Another metric is represented by the closeness centrality. It reflects how "close" a node is to all the others, as it is computed through the reciprocal of the average length of shortest paths between itself and all the other nodes in the network. Lastly, the local clustering coefficient of a node measures the fraction of observed versus possible closed triplets. It reflects how likely are the neighbors of a node connected. A previous study investigated the applicability of different metrics in co-occurrence networks to find keystone species in microbial communities (Berry & Widder, 2014). The investigation showed that keystone species tend to be highly connected centrally-clustering nodes (Berry & Widder, 2014). The results demonstrated that high mean degree, low betweenness centrality, and high closeness centrality can be used to identify keystones with 85% accuracy (Berry & Widder, 2014). However, some non-keystone species also had these properties. This indicates that such properties are a prerequisite but not highly specific for keystone species.

It is not yet understood which measure of node importance best reflects the ecological importance of a species (Röttjers & Faust, 2018). Röttjers & Faust (2018) found that there is a striking lack of overlap in hubs identified through different tools, and the prediction of ecological interactions is hampered by underlying environmental gradients. Moreover, determining hubs is based on associations, but the prediction accuracy of associations is low (Weiss *et al.*, 2016). Finally, is there a way to reliably predict keystone species? This question was addressed in a

Correspondence Letter (Röttjers & Faust, 2019) to the claim that microbial networks can identify keystones (Banerjee *et al.*, 2018). Evidence for accurate keystone prediction from inferred networks is mixed at best (Röttjers & Faust, 2019). Röttjers & Faust (2019) concluded that better network inference tools and more validation experiments are required prior to classifying network hubs as keystones.

Module detection

An ecosystem contains a community of microorganisms that, in turn, may contain subcommunities, i.e., microorganisms that may be responsible for a certain function, and subsequently most likely interact with each other. Detecting groups of microorganisms that are strongly associated transfers to the mathematical problem to find modules within a network. A module M is a subset of the node set V whose members are highly connected to each other, and loosely or not connected to nodes that are not within this module. A module in a network can represent microorganisms that are involved in the same biogeochemical cycle or inhabit the same environmental niche.

Detecting modules is not straightforward and several tools exist. Most of the methods aim to maximize modularity, which quantifies the density of links within modules, in opposition to links between modules (Rahiminejad *et al.*, 2019). Examples of these methods are represented by the Girvan-Newman (Girvan & Newman, 2002), and the Louvain algorithms (Blondel *et al.*, 2008). The limiting factor for applying the Louvain algorithm is the memory requirement rather than the computation time as is the case with Girvan-Newman algorithm (Rahiminejad *et al.*, 2019). The Girvan-Newman algorithm is a network partitioning method that iteratively removes edges. First, it separates all nodes into single-node modules. Next, it recursively combines nodes/modules of the removed edges. The edge with highest betweenness centrality (similar defined to node betweenness centrality) are selected until no edges remain. The betweenness centrality has to be recalculated for those edges affected by an edge removal. In order to select the optimal division among all possible options, modularity is used. The Girvan-Newman algorithm gives good results in many cases, but is impractical for very large networks (Rahiminejad *et al.*, 2019). In contrast, the Louvain algorithm is much faster and detects modules in large networks (Blondel *et al.*, 2008). It contains two steps that are iteratively repeated. First, it separates all nodes into single-node modules. Next, it recursively removes a node from its module and places it in a neighboring one. The gain of modularity is measured and only changes that increase the modularity are kept. The step is repeated for all nodes until a maximum modularity reached. The algorithm is extremely fast and heuristics may further speed it up, e.g., network partitioning stopping when the gain of modularity is under a given threshold (Blondel *et al.*, 2008). Finally, the Girvan-Newman and Louvain algorithms can be extended, allowing to determine modules for multilayer networks

(Didier *et al.*, 2015, 2018).

The partition approaches assign each node to a module, allowing modules of size one. Thus, size threshold may be used to select modules for downstream analysis. One limitation of traditional partition approaches is that they determine disjoint modules, whereas overlapping modules are biologically more relevant. Indeed, the approach of *Overlapping Stochastic Block Model* allows nodes to belong to multiple modules (Latouche *et al.*, 2011). Another drawback of traditional methods is that they do not distinguish between positive and negative associations. A recent biologically-driven algorithm is implemented in manta (Röttjers & Faust, 2020). Contrary to the existing algorithms, manta exploits negative edges while differentiating between weak and strong module assignments to identify biologically relevant modules in real-world data sets (Röttjers & Faust, 2020).

Although, identifying important nodes and modules aids to investigate marine microbial ecosystems, they should be treated with much care and should be validated, if possible, or at least strengthened with in-depth biological knowledge. Evaluating inferred associations is challenging due to the lack of unknown interactions (Bjorbækmo *et al.*, 2019) and microbial associations predict true ecological interactions in a minority of cases (Weiss *et al.*, 2016). Therefore, emergent properties may be more reliable in the search for new biological insights, and microbial association networks provide an excellent tool for studying them (Röttjers & Faust, 2018).

Final remarks

- ⇒ Networks are a great tool to handle the vast number of microorganisms and potential interactions.
- ⇒ Studies employing networks contribute to our understanding of marine microbial interactions.
- ⇒ Network construction is not straightforward and there is a vast collection of tools.
- ⇒ There is no gold-standard network construction tool.
- ⇒ Some tools are better suited than others for the specificities of a dataset.
- ⇒ Identifying important nodes and modules aids to investigate marine microbial ecosystems but should be treated with much care.
- ⇒ Emergent properties may be more reliable in the search for new biological insights.
- ⇒ Microbial association networks allow studying emergent properties.

Chapter 4 Challenges studying microbial interactions

The smallest living organisms

Human relationships may be inferred via connections on social media. Another opportunity to determine professional interactions is to connect researchers who published together. Relationships between macro-organisms, e.g., animals living in the forest, may be observed with bare eyes or binoculars. Unfortunately, microorganism do not maintain a social media account and we cannot observe all microorganisms with our eyes. Even when using microscopes, we cannot easily detect all the tiny microorganisms.

Considering the body size, aquatic microorganisms are traditionally separated into three logarithmic classes, namely pico (0.2–2 μm), nano (2–20 μm), and micro (20–200 μm) (Massana & Logares, 2013). To put it in context, one micrometer corresponds to 10^{-6} meter, i.e., one million micrometer is one meter. If we consider a 0.5 μm microorganism as a 1.70 m human being, a 2 μm microorganism would be four times larger (6.8 m), about the height of a two-story house; and a 20 μm microorganism would be forty times larger (68 m), taller than The Arc de Triomphe (49 m) in Paris.

The invention of microscopes, culturing experiments, and omics technologies allowed to elucidate aspects of the microbial world that were completely hidden. Although the microbial size may not limit any longer their detection, this is not the only challenge on the quest to identify microbial interactions.

Studying microbial interactions experimentally

Most microorganisms remain uncultured and poorly characterized (Baldauf, 2008; Lewis *et al.*, 2020). In their review, Lewis et al. list six factors that influence microbial cultivability (Lewis *et al.*, 2020):

- First, we need to identify the microbial needs regarding substrates and growth conditions.
- Second, we have to learn to induce the transition from the dormant to the active state in persisting microorganisms. They may have evolved different mechanisms to regulate dormancy necessitating different solutions to “resuscitate” them.
- Third, interdependencies between two or more organisms require the identification of partners. Therefore, we need to perform co-isolation and co-culture of partner microorganisms, or to abiotically replace essential partners by providing the substrates the that they would have provided.
- Fourth, we need to identify if partners require a physical contact or specific spatial proximity.

- Fifth, we need to provide essential physico-chemical environmental conditions, e.g., temperature and salinity.
- Sixth, we need to overcome the challenge of culturing low abundant (rare) microorganisms and of competition. A fast-growing microorganism could quickly outcompete a slow-growing target microorganism. Also, a microorganism with a high affinity for an essential limited growth substrate prevents a target microorganism with low substrate affinity to grow. Thus, even if the target microorganism is initially enriched in a sample, its relative abundance can soon be diminished when co-inoculated with competitors. In addition, slow-growing microorganisms may lead to increased time-scales of research, resulting in increased costs.

Although most microorganisms remain uncultured, we can circumvent this problem by identifying and quantifying them through omics technologies, e.g., sequencing the genomes. The genomic revolution started in 1995 with the complete sequencing of the free-living *Haemophilus influenzae* genome, the first published bacterial genome (Fleischmann *et al.*, 1995). Within five years, numerous other bacteria were sequenced. Since then, the genomes of eukaryotic microorganisms have also been sequenced. This led to the massive expansion of sequence data (Hall, 2007). Within few years, the sequencing method industrialized, and the vast majority of sequence data has been and is now generated by large genome centers (Hall, 2007).

Finally, thanks to omics technologies, microbial interactions may be inferred by identifying metabolic dependencies. A metabolic dependency refers to need of an organism of a substrate produced by another one. Genomic data, in combination with transcriptomic and proteomic data, can be used to infer such dependencies. However, because of the huge number of microorganisms, it may not be possible to obtain the complete omics-data of each one.

Quantifying microorganisms

In order to identify a microorganism, we often use a specific sequence, and not the whole genome. Such sequence should be diverse enough to identify microorganisms and conserved enough to compare them. Each organism contains ribosomes and their genetic sequence is well preserved. The 16S rRNA and the 18S rRNA gene sequences can be used for the so-called marker- or targeted-sequencing, for prokaryotes (bacteria and archaea) and eukaryotes, respectively.

These culture-independent methods based on targeted sequencing of ribosomal genes to identify microorganisms date back to the mid-1980s (Olsen *et al.*, 1986). Specifically, the targeted sequencing of environmental 16S rRNA gene revealed the tremendous number of uncultured taxa. Comparing plate counts with direct microscopic counts, isolated cells are estimated to constitute less than 1% of microbial species (Staley & Konopka, 1985). The so called ‘great plate count anomaly’ (Staley & Konopka, 1985) refers to the discrepancy between microorganisms that are

present in a given environment, and those that can be cultured in the laboratory (Lewis *et al.*, 2020).

We know that microorganisms dominate our world, but how many are there? The number of microbial species on Earth is estimated to be $\approx 10^{12}$ (Locey & Lennon, 2016), comprising $\approx 10^{30}$ cells (Whitman *et al.*, 1998; Kallmeyer *et al.*, 2012). The oceans are estimated to harbor 10^{29} microbial cells (Whitman *et al.*, 1998) accounting for ~70% of the total marine biomass (Bar-On *et al.*, 2018).

Sequencing 16S and 18S rRNA genes allows to identify and quantify an enormous number of microorganisms in a given sample. A set of samples can then be used to determine microbial associations. These associations can be considered as edges and microorganisms as nodes in a network. Finally, the analysis of microbial association network is a great tool to investigate marine microbial ecosystems. Yet, there are challenges inferring association networks from microbial sequence abundance data.

Technical challenges inferring association networks

Constructing microbial association networks from sequence abundance data presents three main technical challenges.

The compositional effect

First, microbial sequence abundance data suffer from the compositional effect, which is not an exception but the rule (Gloor *et al.*, 2017). The compositional effect is due to the fixed capacity of the sequencing instruments.

Traditionally, to solve this problem, datasets have been rarified. Shortly, rarefication randomly subsamples each sample until a predefined number of total reads is reached. The probability to choose a read corresponding to a specific ASV is proportional to its sequence abundance. Thus, original proportions are preserved. However, rarefication is not advisable according to Gloor *et al.* (Gloor *et al.*, 2017), because the subsampling results in loss of information and precision (McMurdie & Holmes, 2014). However, there may be already a loss of information and precision because of the high technical variability of rRNA gene sequencing (Faust, 2021).

Another solution is to ratio transform the data. Ratio transformation captures the relationships between the microorganisms (Gloor *et al.*, 2017). Applying the logarithm (log-ratios) results in symmetric and linearly related data. We can obtain information about the log-ratio of microbial sequence abundances relative to other microorganisms, and this information is directly relatable to the environment (Gloor *et al.*, 2017). A popular method is the centered-log-ratio (clr) transformation introduced by Aitchison (Aitchison, 1986). As log-ratios are sensitive

to zeros leading to negative infinities, a pseudo-count is usually added. Alternatives exists, but they make additional assumptions about zeros that have not yet been validated (Martín-Fernández *et al.*, 2012; Tsilimigras & Fodor, 2016).

Sparseness of data

Another challenge is represented by the sparseness of data. Data is sparse if the sequence abundance table contains many zeros, which means the microorganisms (ASVs) were absent in most samples. A zero could mean the absence of a microorganism (structural or essential zero), or that the microorganism has not been detected, i.e., undersampled (rounded zero) (Martín-Fernández *et al.*, 2003). In the latter case, the microorganism was present, but we failed to detect it, for example if by chance rare components are not present in the sample drawn from the microbial community (Friedman & Alm, 2012). Alternatively, microorganisms may not be detected due to an insufficient sequencing depth or the microbial sequence was not amplified during the PCR.

Sparse data can cause artefactual associations for low-abundance microorganisms with very few non-zero observations (Aitchison, 1981). Thus, removing extremely rare microorganisms prior to network construction subsequently lowers artefactual associations. There are two filtering solutions, both using an arbitrary threshold (Faust, 2021): i) using a prevalence filter that removes microorganisms appearing in too few samples; and ii) not considering (computing) associations between pairs that have a large number of matching zeros. Choosing the threshold is not straightforward and different formulas have been proposed (Cougoul *et al.*, 2019). Furthermore, filtering could significantly reduce the computational time needed to infer networks and the quality of network inference (Cougoul *et al.*, 2019). The heterogeneous mode implemented in the network construction tool FlashWeave ignores matching zeros when computing associations (Tackmann *et al.*, 2019).

Small observation-to-variables ratio

Traditionally, statistical analysis requires more observations than variables. However, most datasets display a small observations-to-variables ratio, i.e., the ratio of the number of samples (observations) and the number of components (variables) is low and often with more detected microorganisms (ASVs) than the number of samples. Obtaining more observations (more samples) is limited because of costs, time, and resources. Several network construction tools aim to be robust to the small observation-to-variables ratio and also the compositional effect (Layeghifard *et al.*, 2017), e.g. SPIEC-EASI (Kurtz *et al.*, 2015) and CoNet (Faust *et al.*, 2012). Moreover, removing rare microorganisms (e.g., using a prevalence filter) when controlling for data sparsity, will also reduce the number of variables.

From association networks to biological meaningful interpretations

In addition to technical challenges in the construction of microbial association networks, there are three main interpretation challenges.

Ecological networks do not equal association networks

First, 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: negative interactions like parasitism or competition; positive interactions like symbiosis; and neutral interactions (Xiao *et al.*, 2017). On the contrary, positive associations represent high niche overlap and/or positive microbial interactions, and negative associations indicate divergent niches and/or negative microbial interactions (Hernandez *et al.*, 2021). Moreover, ecological networks are directed graphs, where the edges point from a start node (source) to an end node (target). In contrast, association networks are (usually) undirected. Although association networks provide ecological insight, they do not encode casual relationships or observed ecological interactions. Thus, unless edges are verified with experiments or additional information, one should be careful when attributing biological meaning to network properties (Röttjers & Faust, 2018).

Dense networks

One major problem after network construction is represented by the presence of too many inferred edges resulting in dense networks, also called hairball networks (Röttjers & Faust, 2018). The large number of edges may result in a poorly informative network and weaken the biological interpretation. Less dense networks may be obtained when lowering the commonly used corrected p-value of 0.05 to 0.001 (Weiss *et al.*, 2016), and increasing the cut-off for other criteria, such as the association strength, prevalence, and abundance filtering (Röttjers & Faust, 2018). Another strategy to circumvent dense networks is agglomeration through taxonomic or ecological (functional) groupings (Lima-Mendez *et al.*, 2015). However, depending on the research aim, such grouping may not be applicable or desired.

Indirect dependencies

For most microbial association networks, an edge could either indicate ecological interaction or a similar/different environmental preference. This is an effect of indirect dependencies. Specifically, for most microbial association networks, an edge indicates: i) a true ecological interaction between two microorganisms; ii) similar or contrary dependence to environmental factor/s or a third microorganism (indirect edge); or iii) an association by chance. The latter two options do not predict microbial interactions. Indirect edges occur when two microorganisms are

both dependent on an abiotic environmental factor (e.g., same nutrients and temperature requirements) or a biotic factor (e.g., same prey or predator), but do not interact with one another. Thus, it is necessary to disentangle the nature of a given association in a network by determining whether or not it is environmentally-driven.

In order to use the network as a representation of the microbial ecosystem, environmentally-driven associations need to be removed before analysis and interpretation. However, there are too many potential associations in an inferred network to experimentally test each potential interaction. More importantly, most microorganisms are uncultured and, subsequently, not available for experimental testing in the lab. We list and describe our and other existing approaches to disentangle environmentally-driven associations in Chapter 5.

Comparing networks

Differences in the manner a network is obtained

There are common and well-known challenges for comparing networks that were generated with different strategies. First, the measurement (e.g., sequencing depth), the spatial and temporal frequency of the sampling, and the availability of replicates can influence network construction (Faust *et al.*, 2015). Second, each step from the samples to the sequence abundance data may introduce a bias, e.g., when obtaining different amounts of DNA extracted, amounts sequenced, and percentages of high-quality reads (Faust & Raes, 2012).

Next, comparing networks obtained through different construction tools is challenging because each of them infers different numbers and types of significant edges for the same data (Weiss *et al.*, 2016). Weiss *et al.* (2016) compared eight different network construction tools demonstrating an average of 31.5% shared edge for all pairwise combinations of tools and all datasets/models tested. Finally, different cut-off levels for association strengths may have been used and alter networks' structure (Connor et al., 2017), further complicating comparisons.

Differences in network properties

The environmental influence on network properties is unclear, and, subsequently, it is unknown if such properties can be reliably inferred from microbial networks (Röttjers & Faust, 2018). Röttjers and Faust (2018) showed that different levels of environmental influence changed network structure and that simulated networks fail to match global properties of the underlying true interaction network. This indicates that it is unwise to attribute biological relevance to the properties without further experiments or additional information (Röttjers & Faust, 2018).

The actual network comparison

Network comparison is an active field of research. There are several ways to compare networks

and measure their similarities, for example networks from different environments, locations, or times. A simple approach considers common and unique nodes and edges. Their number can be used to quantify network (dis)similarity, as proposed in the study of (Poisot *et al.*, 2012). Another approach uses distance probability distributions (Schieber *et al.*, 2017).

We can also compare specific network properties (global metrics), e.g., edge density and transitivity. However, global metrics may disregard local patterns. Two local-topological metrics use either motifs or graphlets. The latter preserve all connections among nodes (induced subgraph), in contrast to the former. Graphlets can be used for network comparison (Pržulj *et al.*, 2004), which allows to compare network topology without considering specific microorganisms. Not all graphlets need to be considered because of redundancy (Yaveroğlu *et al.*, 2014). In addition, instead of using graphlets up to four nodes (Yaveroğlu *et al.*, 2014), we can preferentially choose graphlets up to three nodes (Espejo *et al.*, 2020) when comparing (many) large networks.

However, microbial association networks contain a large number of errors and derived properties do not necessarily reflect the true community structure (Röttjers *et al.*, 2020). Röttjers *et al.* (2020) developed a toolbox to investigate noisy networks with null models to identify non-random patterns in groups of association networks. Comparing multiple networks identifies conserved subsets, so called core-association networks, and other properties that are shared across all networks (Röttjers *et al.*, 2020).

Comparing nodes in a network

Various methods exist to compare microorganisms, e.g., using the size, abundance, genomic sequence similarity; determining their seasonality; or using network-based metrics, such as the degree, betweenness, and closeness centrality scores. Moreover, graphlets can also be used for a network-based comparison of microorganisms, by extending the single value of the node degree to the graphlet degree vector quantifying the different nodes' *roles* (i.e., quantifying the orbits a node touches). Similar network-based roles potentially translate to similar functional roles in the ecosystem.

The single static network

Lastly, it may also be problematic that often single static networks are used to represent dynamic microbial ecosystems. The obvious solution would be to construct a network for each location, time point (temporal networks), and environmental condition. For instance, temporal networks are conceptually well-defined within the mathematical field. They have broad applications, e.g., transportation, social, and biological networks. Already a decade ago, Prztycka *et al.* (2010) stated that it is essential for computational biologists studying (molecular) networks to shift their focus

to developing methods that incorporate information about the dynamic nature of (cellular) systems. Temporal networks can have several names, e.g., time-dependent, evolving, time-varying, and historical networks (Wang *et al.*, 2019). However, current marine microbial datasets do not allow the construction of temporal networks since they often only provide one single sample per time point. Similarly, each location is often sampled once in global scale studies. In this thesis, we present our approach to overcome the problem of not having the required sampling size per time point (Chapter 6) or location (Chapter 7).

Final remarks

- ⇒ Studying microbial interactions is challenging.
- ⇒ Besides sampling problems, there are challenges before, during, and after inferring interactions through association networks.
- ⇒ Nevertheless, networks are a valuable tool given their advantages.
- ⇒ Although they do not provide complete information on the interactions inside the system, the network provides a system view, which has value on itself.
- ⇒ Improving inferred networks before down-stream analysis will benefit microbial interaction studies.

Part II Disentangling marine microbial association networks

Chapter 5 Disentangling environmental effects in microbial association networks

Ina Maria Deutschmann, Gipsi Lima-Mendez, Anders K. Krabberød, Jeroen Raes, Sergio M. Vallina, Karoline Faust and Ramiro Logares

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

Introduction

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 microorganisms 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 microorganism 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 (Table 2). 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 3 and Table 3). 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.

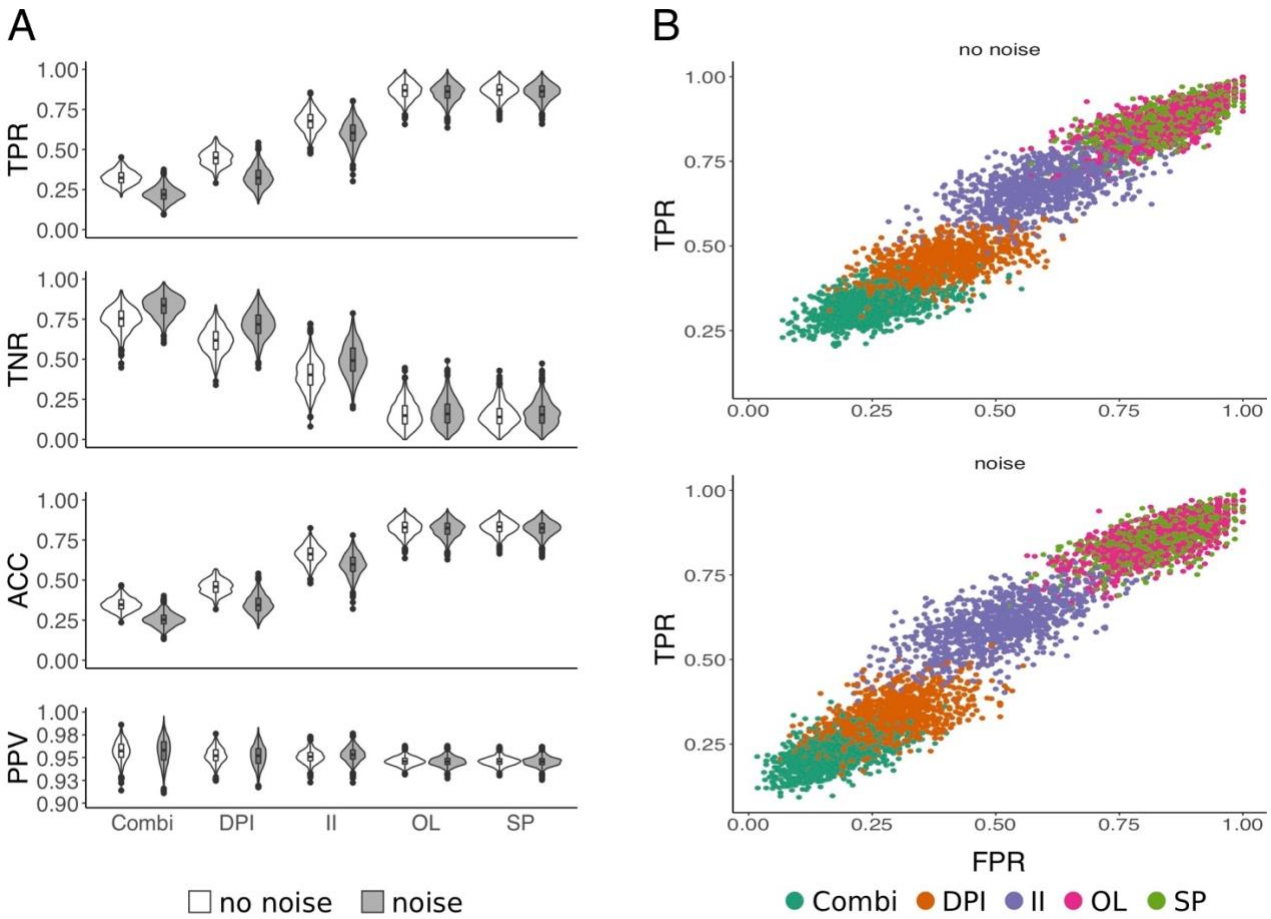


Figure 3: **Evaluation of EnDED: intersection combination and individual methods on simulated networks.** Using 1000 simulated networks, and 1000 simulated networks incorporating noise, we evaluated EnDED's performance. Plot A) displays the evaluation measurements true positive rate (TRP), true negative 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.

Table 2**: 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

Table 3**: 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

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 3). 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 3).

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 4).

We found that 25230 (84.6%) of the network edges were in at least one and in maximum six environmental triplets (Figure 4 and Table 4). 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^-.

Table 4: **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

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 4 and Table 5). 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.

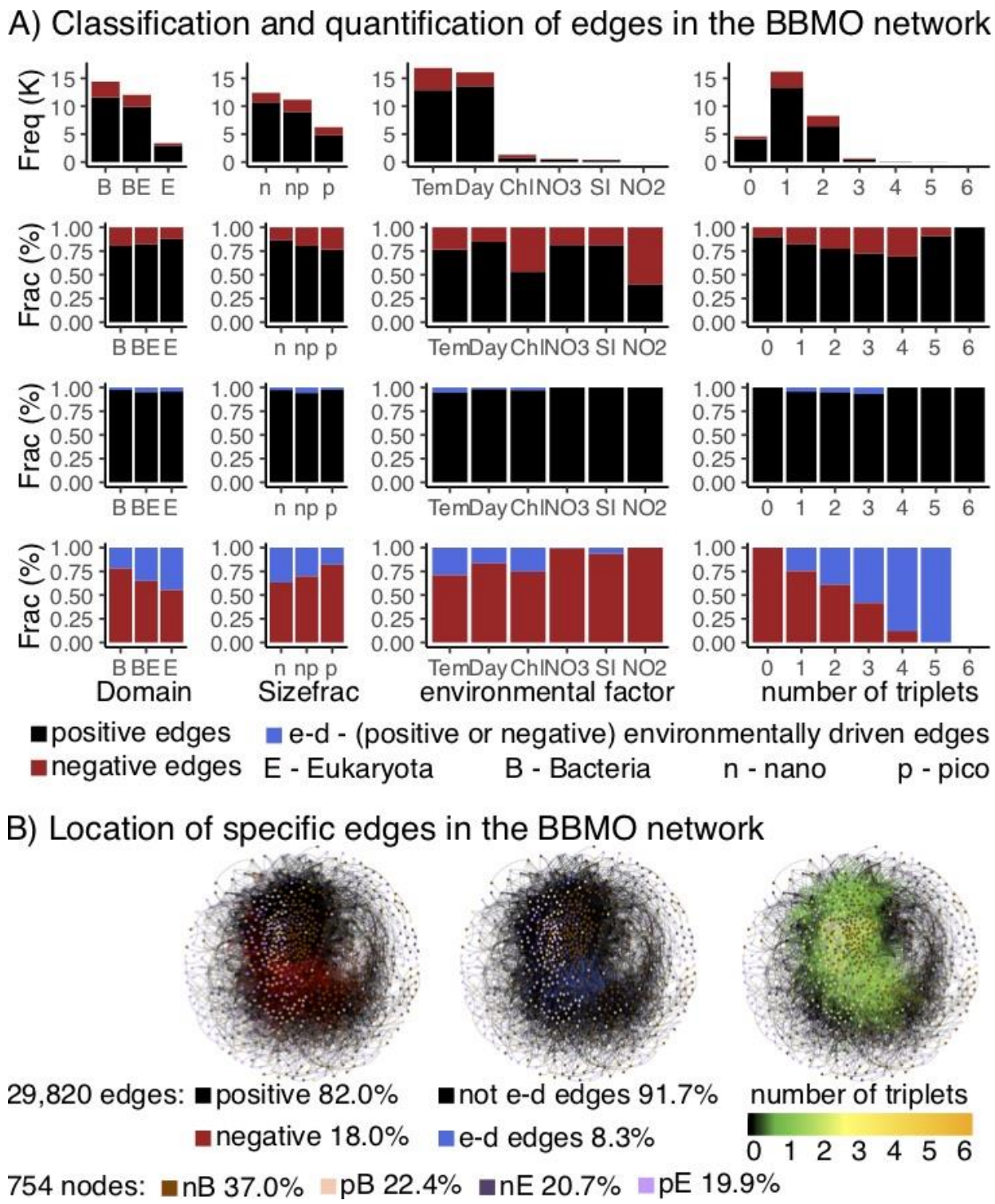


Figure 4: **Quantification of environmentally-driven associations in the BBMO network.** For A) the first column shows the number and fraction of microbial associations divided by domain: Bacteria-Bacteria associations (B), Bacteria-Eukaryote associations (BE), and Eukaryote-Eukaryote associations (E). 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.

Table 5**: 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

We also determined for each association the Jaccard index, which indicates how often two microorganisms appear together in the dataset. We assumed that two microorganisms 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 6). 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.

Table 6: **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%)

The intersection combination removed more negative than positive edges, 1554 and 934, respectively (Figure 4). 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:

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

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 7). 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 7). Ideally none of these described associations should be removed by EnDED. Yet, the intersection combination removed five associations (Table 7). 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.

Table 7**: 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.

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 (2018) simulated an increasing environmental influence and observed a decrease in retrieving true interactions from inferred associations. 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, Zhou, *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, Zhou, *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 chapter, 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[7]. 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:

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

[7] Diagonal coefficients are set to zero using the Klemm-Eguíluz algorithm when generating the modular scale-free matrix but later set to -0.5.

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 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 chlorophyll (µg/l), 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 SAIV-AS-SD204 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 microorganisms. 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 5).

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.

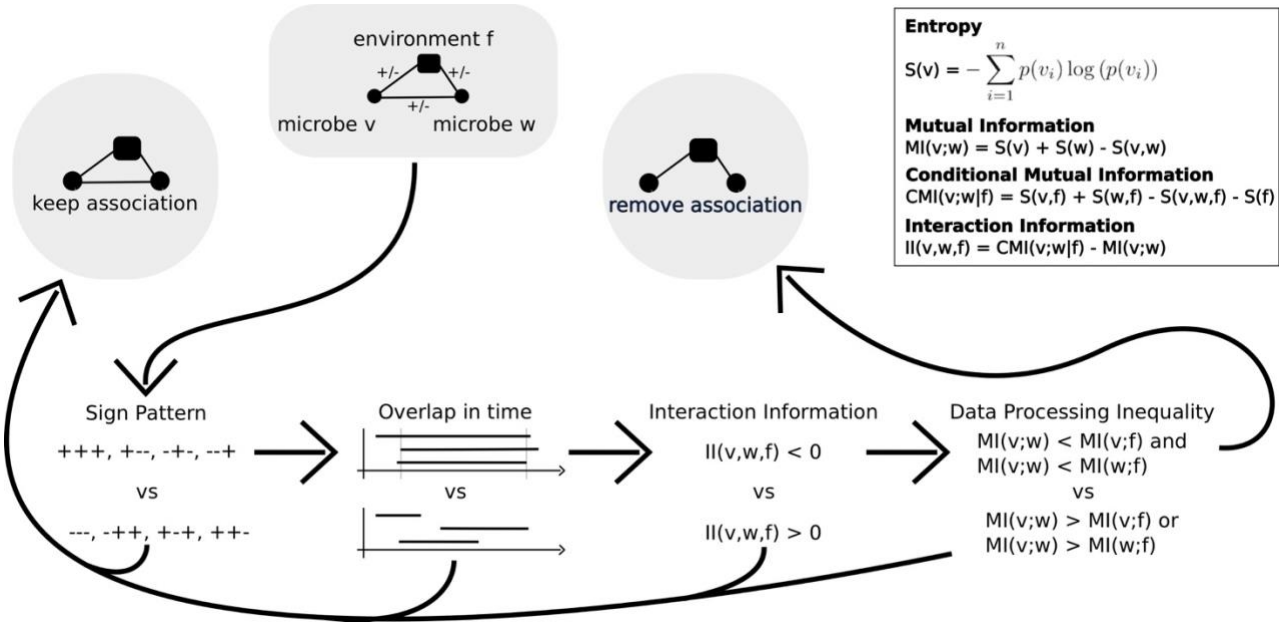


Figure 5: **EnDED Methods Overview.** EnDED is an implementation of four methods aiming to determine whether an edge between two microorganisms is indirect through the action of an environmental factor. The four methods are: Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality (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.

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 5). 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 5, if O>60%, the microbial association is considered environmentally-driven.

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

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

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

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

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

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:

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

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

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

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

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

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

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:

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

In other works (Ghassami & Kiyavash, 2017), the II is defined with a different sign convention: $\mathrm{II} = \mathrm{MI} - \mathrm{CMI}$. 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 $i \in \{1, \ldots, 1000\}$. Afterwards, we quantified with c how many random CMI_i were at least as small as the original CMI_i: $c = |i: \mathrm{CMI}_i \leq \mathrm{CMI}_{original}, i \in \{1, \ldots, 1000\}|$. We calculated the p-value as

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

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, MI($\boldsymbol{v}$; $\boldsymbol{w}$) is smaller than MI($\boldsymbol{v}$; $\boldsymbol{f}$) and 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 $\boldsymbol{v} = (v_1, \ldots, v_n)$ 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: $P = TP + FN$, and N as the number of all true interactions, i.e., all true negative and false positive detected environmentally-driven edges: $N = TN + FP$. Then, we calculated the true positive rate (sensitivity), by dividing the number of true positives by the number of all real positives: $TPR = (TP)/(P)$. Equivalently, we can also calculate the true negative rate (specificity) by dividing the number of true negatives by the number of all real negatives, $TNR = (TN)/(N)$. The false positive rate (fall out) is the complementary to TNR, i.e. $FPR = 1 - TNR$. The positive predictive value (precision) can be calculated by dividing the number of true positives by the sum of all predicted positives, $PPV = (TP)/(TP + FP)$. The accuracy is calculated by dividing the sum of true positives and true negatives by the sum of all real positives and real negatives, $ACC = (TP + TN)/(P + N)$.

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.

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 commands 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.

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).

Final Remarks

- ⇒ Associations could result from either ecological interactions between microorganisms, or environmental selection.
- ⇒ Determining a cut-off level for the association score is not sufficient to separate true from false interactions (simulated networks).
- ⇒ EnDED is an implementation of four approaches and their combination (a newly developed ensemble approach) to predict environmentally-driven microbial associations.
- ⇒ EnDED can be used to quantify environmentally-driven associations for each environmental factor allowing to determine the main environmental drivers of indirect associations.

⇒ The main drivers of environmentally-driven edges in the real data BBMO network based on ten years of data at the BBMO were temperature and day length, to a lesser extent total chlorophyll and nutrients.

⇒ The fraction of environmentally-driven edges among negative microbial associations increased rapidly with the number of environmental factors (real network).

⇒ EnDED should be included in a filtering strategy to reduce the number of false associations and consequently the number of potential interaction hypotheses.

⇒ Proposed idea: quantify edge recurrence (temporal and spatial).

⇒ Observation: most of the edges are within one environmental triplet, some within no triplet and EnDED tests for indirect dependencies only within a triplet and the corresponding environmental factor.

Chapter 6 Disentangling temporal associations in marine microbial networks

Ina Maria Deutschmann, Anders K. Krabberød, L. Felipe Benites, Francisco Latorre, Erwan Delage, Celia Marrasè, Vanessa Balagué, Josep M. Gasol, Ramon Massana, Damien Eveillard, Samuel Chaffron and Ramiro Logares

Abstract

Microbial interactions are fundamental for Earth's ecosystem functioning and biogeochemical cycling. Nevertheless, they are challenging to identify and remain barely known. The omics-based censuses are helpful to predict microbial interactions through the inference of static association networks. However, since microbial interactions are highly dynamic, we have developed a post-network-construction approach to generate a temporal network from a single static network. We applied the approach to understand the monthly microbial associations' dynamics occurring over ten years in the Blanes Bay Microbial Observatory (Mediterranean Sea). For the decade, we identified persistent, seasonal, and temporary microbial associations. Moreover, we found that the temporal network appears to follow an annual cycle, collapsing and reassembling when transiting between colder and warmer waters. We observed higher repeatability in colder than warmer months. Altogether, our results indicate that marine microbial networks follow recurrent temporal dynamics, which need to be accounted to better understand the dynamics of the ocean microbiome.

Keywords: association network; temporal network; time series; microbial interactions; permanent versus temporary associations; network collapse and reassembling; bacteria and micro-eukaryotes; Mediterranean Sea

Introduction

Microorganisms are the most abundant life forms on Earth and are fundamental for global ecosystem functioning (Falkowski *et al.*, 2008; DeLong, 2009; Krabberød *et al.*, 2017). The number of microorganisms on Earth is estimated to be $\approx 10^{12}$ species (Locey & Lennon, 2016), comprising $\approx 10^{30}$ cells (Whitman *et al.*, 1998; Kallmeyer *et al.*, 2012). The oceans harbor $\approx 10^{29}$ microbial cells (Whitman *et al.*, 1998) accounting for ~70% of the total marine biomass (Bar-On *et al.*, 2018; Bar-On & Milo, 2019). These cell numbers are known to be dynamic.

Microbial ecosystems are dynamic and their community composition is determined through a combination of ecological processes: selection, dispersal, drift, and speciation (Vellend, 2020). Selection is a prominent community structuring force that is exerted via multiple abiotic and biotic environmental factors (Lindström & Langenheder, 2012; Mori *et al.*, 2018). Several studies have addressed the role of *abiotic* factors in structuring microbial communities. For example, temperature, one of the primary environmental variables, exerts selection in the ocean microbiome over spatiotemporal scales (Bunse & Pinhassi, 2017; Giner *et al.*, 2019; Lambert *et al.*, 2019; Logares *et al.*, 2020). *Biotic* factors can also exert a strong selection on microbial communities (Barraclough, 2015). However, a mechanistic understanding of how they affect community

structure is currently lacking, as the diversity of microbial interactions is barely known (Krabberød *et al.*, 2017; Bjorbækmo *et al.*, 2019).

The vast microbial diversity and the fact that most microorganisms are still uncultured (Baldauf, 2008; Lewis *et al.*, 2020) make it impossible to experimentally test all potential interactions. However, omics-technologies allow estimating microbial sequence abundances over spatiotemporal scales, which permit determining (statistical) associations between microorganisms. These associations can be summarized as a network with nodes representing microorganisms and edges representing potential interactions (Weiss *et al.*, 2016; Layeghifard *et al.*, 2017).

As microorganisms are highly interconnected (Layeghifard *et al.*, 2017), association networks provide a general overview of the entire microbial system and have been tremendously valuable for generating interaction hypotheses. In particular, several time-series have allowed the investigation of possible ecological interactions among marine microorganisms (Steele *et al.*, 2011; Chow *et al.*, 2013, 2014; Cram, Xia, *et al.*, 2015; Needham *et al.*, 2017; Parada & Fuhrman, 2017; Krabberød *et al.*, 2021). For example, previous work characterized ecological links between marine archaea, bacteria, and eukaryotes (Steele *et al.*, 2011), including links with viruses (Chow *et al.*, 2014; Needham *et al.*, 2017), also investigating within- and between ocean-depth relationships (Cram, Xia, *et al.*, 2015; Parada & Fuhrman, 2017). Not only time-dependent associations among ecologically important taxa were identified, but also potential synergistic or antagonistic relationships, as well as possible 'keystone' species and potential niches (Steele *et al.*, 2011; Chow *et al.*, 2013). Moreover, studies found more associations between microorganisms than between the microorganisms and environmental factors, which would suggest the dominance of microbial relationships over associations between microorganisms and environmental factors (Steele *et al.*, 2011; Krabberød *et al.*, 2021).

Previous studies used temporal microbial-abundance data to build static networks. This static abstraction is based on several assumptions (Blonder *et al.*, 2012), principally that the network topology does not change (static) and edges represent persistent associations assumed as interactions, that is, edges are present throughout time-space. This assumption cannot represent the reality for most microbial interactions. Thus, a single static network usually captures persistent, temporary, and recurring (including seasonal) associations, which need to be disentangled.

Despite the contribution of static networks to our understanding of microbial interactions in the ocean, it is necessary to incorporate the temporal dimension. Using a temporal network instead of a single static network would allow investigating the dynamic nature of microbial associations, contributing to comprehend how they change over time, whether their change is deterministic or stochastic, and how environmental selection influences network architecture. Addressing these questions is fundamental for a better understanding of the dynamic interactions that underpin microbial ecosystem function. Here, we investigate marine microbial associations through time using an approach developed to determine a temporal network from a single static network.

Results

Extracting a temporal network from a single static association network

Leveraging ten years of monthly samples from the Blanes Bay Microbial Observatory (BBMO) in the Mediterranean Sea (Gasol *et al.*, 2016), we computed sequence abundances for 488 bacteria and 1005 microbial eukaryotes from two organismal size-fractions: picoplankton (0.2 – 3 µm) and nanoplankton (3 – 20 µm). We removed Archaea since they are not very abundant in the BBMO surface and, additionally, primers were not optimal to quantify them. We inferred Amplicon Sequence Variants (ASVs) using the 16S and 18S rRNA-gene. After filtering the initial ASV table for sequence abundance and shared taxa among size fractions, we kept 285 and 417 bacterial, 526 and 481 eukaryotic ASVs in the pico- and nanoplankton size-fractions, respectively. We found 214 bacterial ASVs that appeared in both size fractions, but only two eukaryotic ASVs: a *Cryothecomonas* (Cercozoa) and a dinoflagellate (Alveolate).

We used a total of 1709 ASVs to infer a preliminary association network with the tool eLSA (Xia *et al.*, 2011, 2013). Next, we removed environmentally-driven edges with EnDED (Deutschmann *et al.*, 2020) and only considered edges which association partners co-occurred more than half of the times together than alone (see methods and Figure 6A-B). Our filtering strategy removed a higher fraction of negative than positive edges (see methods and Table 8). The resulting network is our single static network connecting 709 nodes via 16626 edges (16481 edges, 99.1%, positive and 145, 0.9% negative).

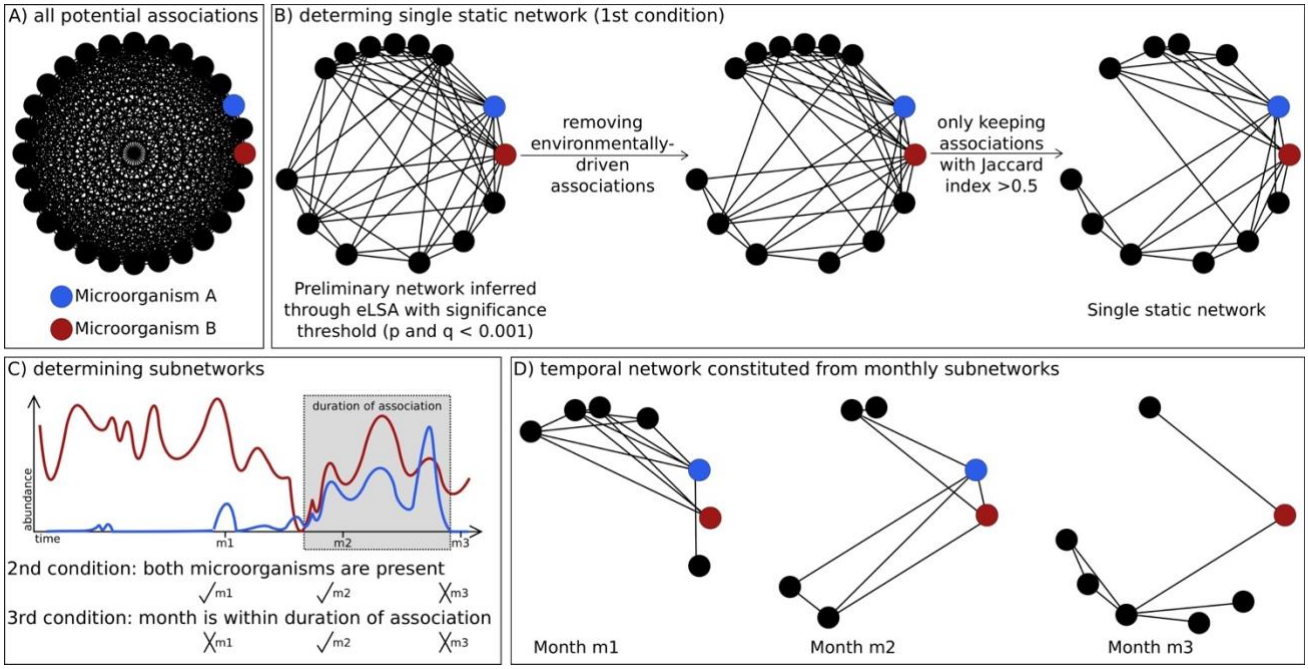


Figure 6: **Conceptual idea on how we determine a temporal network from a single static network via subnetworks**. A) A complete network would contain all possible associations (edges) between microorganisms (nodes). B) The single static network is inferred with the network construction tool eLSA and a filtering strategy considering association significance, the removal of environmentally-driven associations, and associations whose partners appeared in more samples together than alone, i.e., Jaccard index being above 0.5. An association having to be present in the single static network is the first out of three conditions for an association to be present in a monthly subnetwork. C) In order to determine monthly subnetworks, we established two further conditions for each edge. First, both microorganisms need to be present in the sample taken in the specific month. Second, the month lays within the time window of the association inferred through the network construction tool. Here, three months are indicated as an example. D) Example of monthly subnetworks for the three months. The colored nodes correspond to the abundances depicted in C).

Next, we developed a post-network-construction approach to determine a temporal network from a single static network. Building upon the single static network, we determined 120 sample-specific (monthly) subnetworks (see methods for details). These monthly subnetworks represent the 120 months of the time series and together comprise the temporal network. Each monthly subnetwork contains a subset of the nodes and a subset of the edges of the single static network. To determine which nodes and edges are present each month, we used the ASV abundances indicating the presence (ASV abundance > 0) or absence (ASV abundance = 0) as well as the estimated start and duration of associations inferred with the network construction tool eLSA (Xia *et al.*, 2011, 2013) (Figure 6, see Methods).

Table 8: **Number of nodes and edges in preliminary networks and the temporal network.** Number of nodes, removed isolated nodes, and number and fraction of edges in the preliminary network (A), and network obtained after removing environmentally-driven edges (B) and edges with association partners appearing more often alone than with the partner (C), which is the single static network. For comparison, we also give the minimum and maximum number of nodes and edges for the temporal network (D). We did not determine the union and intersection for the temporal network. If an ASV appeared in the nano and pico size fraction, it is counted twice. Therefore, for A-C) we also determined the number of microorganisms not considering size fraction (union) and being present in both size fractions (both, i.e. intersection).

	A) eLSA	B) EnDED	C) Static network	D) Range in Temporal network
Connected nodes	754	754	709	130-542
Bacteria (pico)	169	169	164	13-148
Bacteria (nano)	279	279	251	31-204
Bacteria (union)	309	309	281	
Bacteria (both)	139	139	134	
Eukaryote (pico)	150	150	141	7-124
Eukaryote (nano)	156	156	153	2-138
Eukaryote (union)	306	306	294	
Eukaryote (both)	0	0	0	
Isolated nodes	1000	0	45	6-38
Edges	29820	26505	16626	538-15083
Positive edges	24458	23405	16481	523-14940
(%)	82.0	88.3	99.1	92.2-99.7
Negative edges	5362	3100	145	12-143
(%)	18.0	11.7	0.9	0.3-7.8

pico and nano – microorganisms detected in the picoplankton and nanoplankton, respectively, union – how many microorganisms when not considering size-fraction, both – how many microorganisms appear in both size fractions

The single static network metrics differed from most monthly subnetworks

Since each monthly subnetwork was derived from the single static network, they were smaller, containing between 141 (August 2005) and 571 (January 2012) nodes, median ≈354 (Figure 7A), and between 560 (April 2006) to 15704 (January 2012) edges, median ≈6052 (Figure 7B). For further characterization, we computed six global network metrics (Figure C and Methods). The results indicated that the single static network differed from most monthly subnetworks and it also differed from the average. In general, the single static network was less connected (edge density) and more clustered (transitivity) with higher distances between nodes (average path length) and stronger associations (average positive association score) than most monthly subnetworks (Figure 7C). In addition, the single static network was usually more assortative according to the node degree but less assortative according to the domain (bacteria vs. eukaryote) than most monthly subnetworks (Figure 7C). High assortativity indicates that nodes tend to connect to nodes of similar degree and domain, respectively.

Monthly subnetworks display seasonal behavior with yearly periodicity

Over the analyzed decade, the network became more connected and clustered in colder months, with stronger associations and shorter distances between nodes (Figure 7C, Figure 8, and Figure 9). Most global network metrics indicated seasonal behavior with yearly periodicity (Figure 7C). For instance, edge density, average positive association score, and transitivity were highest at the beginning and end of each year, while average path length and assortativity (bacteria vs. eukaryotes) were highest in the middle of each year. Assortativity (degree), in contrast to other metrics, usually had two peaks per year corresponding to April or May, and November (Figure 7C).

We found that mainly temperature and day length, and to a lesser extent nutrient concentrations (mainly SiO_2, NO_3^- and NO_2^-, less PO_4^{3-}), and total chlorophyll-a concentration affected network topologies as indicated by correlation analysis (Figure 9). For example, edge density was highest and temperature lowest in January-March. Then, the edge density dropped as the temperature increased. April-June displayed edge densities slightly above or similar to the warmest months July-September, while October-December had similar or slightly lower edge densities than the coldest months January-March. Edge density vs. hours of light (day length) indicated a yearly recurrent circular pattern for September-April (Figure 8). May-August were not part of the circular pattern and had the highest day length and lowest edge density (Figure 8).

Next, we quantified how many edges are preserved (kept), lost, and gained (new) in consecutive months. We found the highest loss of edges in April. The overall number of edges (preserved and gained) were lowest during April-September and increased towards the end of each year (Figure 7B). The number of associations changed over time in a yearly recurring pattern with few associations being preserved when transitioning from colder to warmer waters. We see a clear network change from colder to warmer months, similar to a crash. In turn, the network change from warmer to colder months is less abrupt, similar to a reassembling. Thus, network change was not symmetrical over the studied decade at BBMO. Moreover, we defined summer and winter as in (Krabberød *et al.*, 2021), and compared both seasons between consecutive years in terms of preserved, gained and lost associations and ASVs. We observed higher repeatability in terms of edges (Figure 10) and ASVs (results not shown) in colder than in warmer months, indicating higher predictability during low temperature seasons.

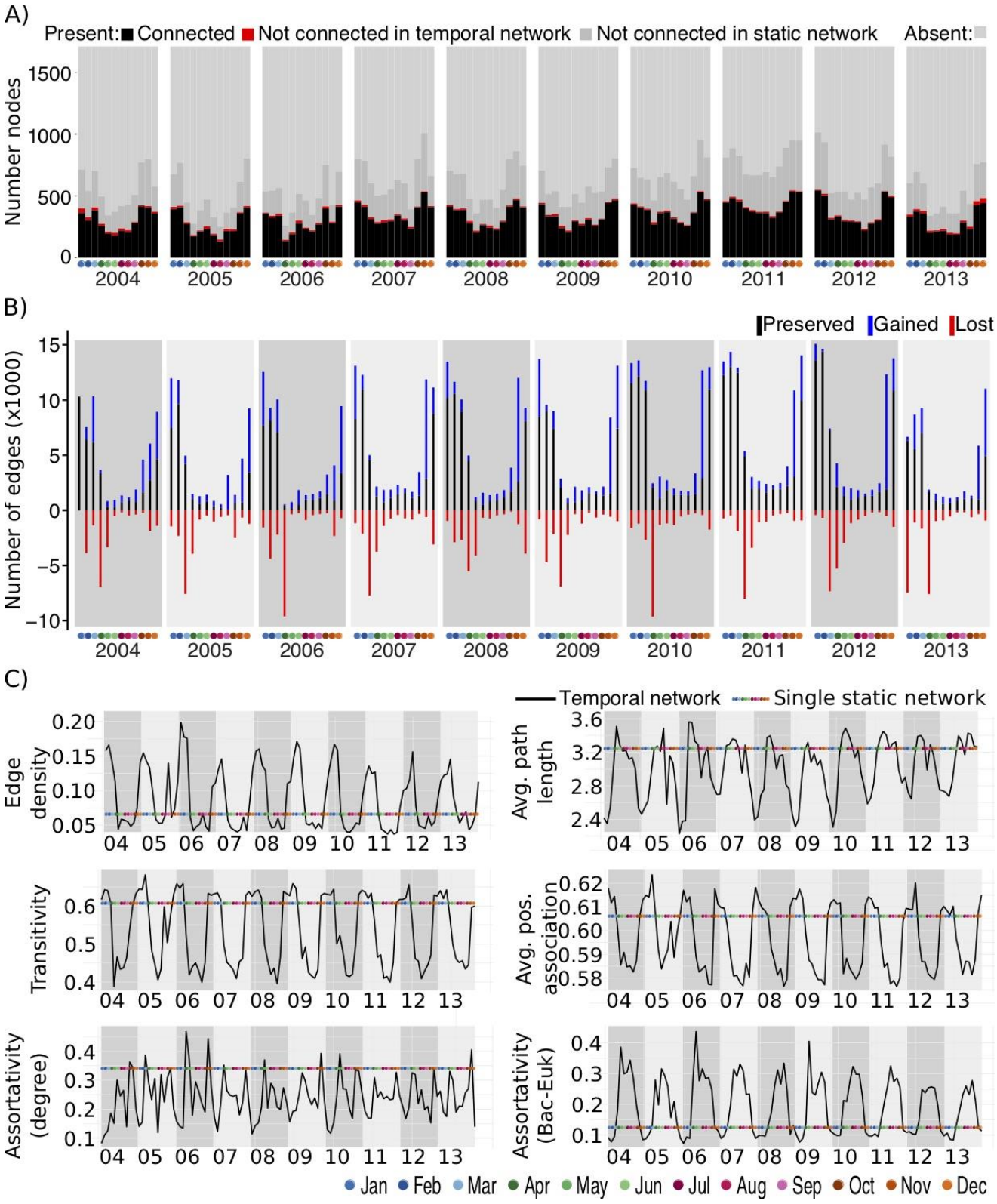


Figure 7: **Global (sub)network metrics**. A) Number of ASVs (counting an ASV twice if it appears in both size fractions) for each of the 120 months of the Blanes Bay Microbial Observatory time series. There are 1709 ASVs, of which 709 ASVs are connected in the static network. In black, we show the number of nodes connected in the temporal network, and in red the number of nodes that are isolated in the temporal network, i.e., they are connected in the static network and have a sequence abundance above zero for that month ("non-zero"). In dark gray, we show the number of ASVs that are non-zero in a given month but were not connected in the static and subsequently temporal network. In light gray, we show the number of ASVs with zero-abundance in a given month. The sum of connected and isolated nodes and non-zero ASVs represents each month's richness (i.e., number of ASVs). B) By comparing the edges of two consecutive months, i.e., two consecutive monthly subnetworks, we indicate the number of edges that have been lost (red), preserved (black), and those that are gained (blue), compared to the previous month. C) Six selected global network metrics for each sample-specific subnetwork of the temporal network. The colored line indicates the corresponding metric for the single static network.

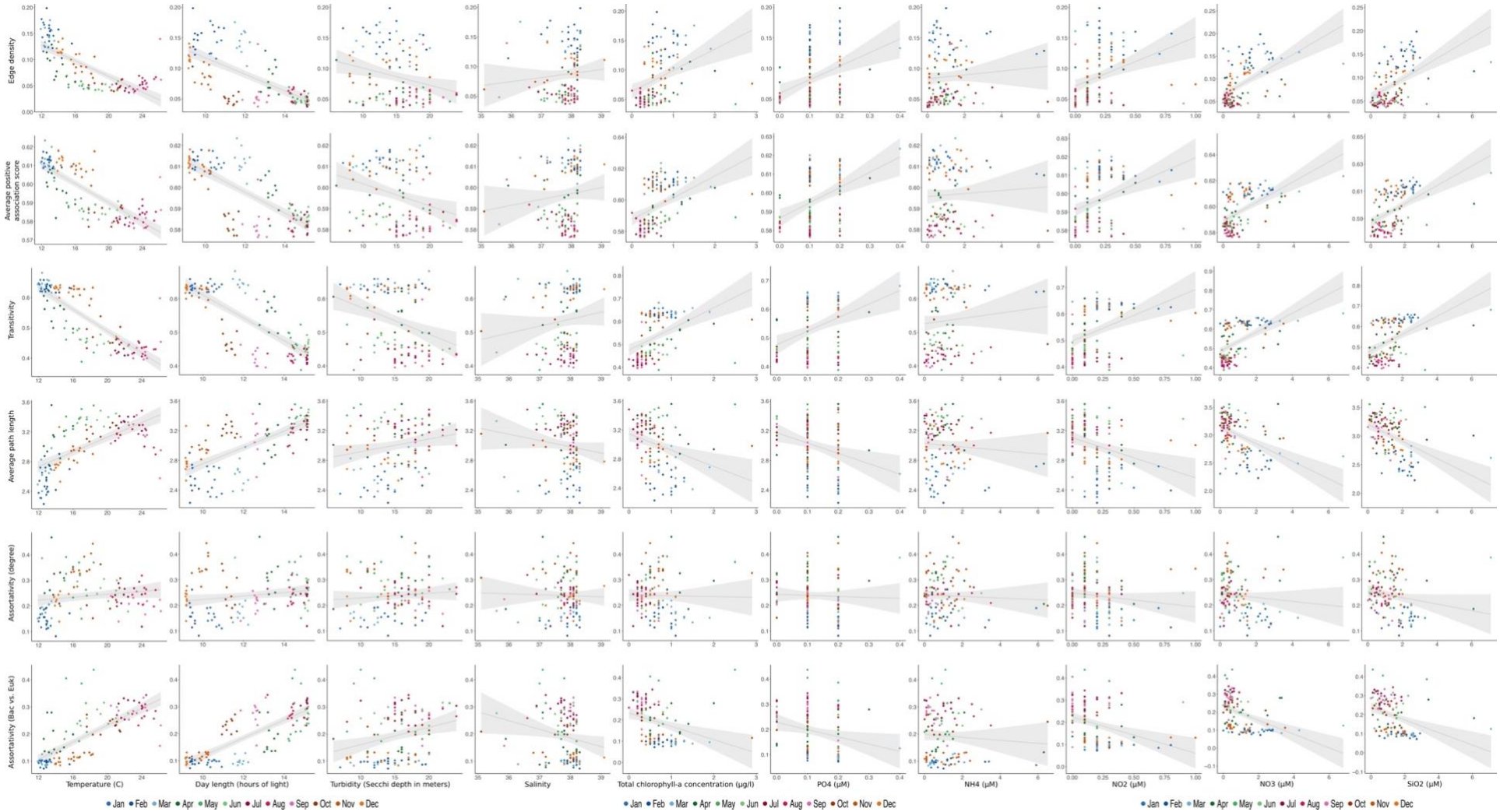


Figure 8**: Correlation Analysis**. Using the temporal network, we correlated six global network metrics with environmental factors including the nutrients PO_4^{3-}, NH_4^+, NO_2^-, NO_3^- and SiO_2. The global network metrics were: Edge density, Average positive association (Avg. pos. ass.) score, Transitivity, Average path length (Avg. path length), Assortativity (degree), and Assortativity (bacteria vs. eukaryote). Each dot is a sample-specific subnetwork and its color indicates the month it represents. Also, the linear regression line with a 0.95 confidence interval is shown in gray.

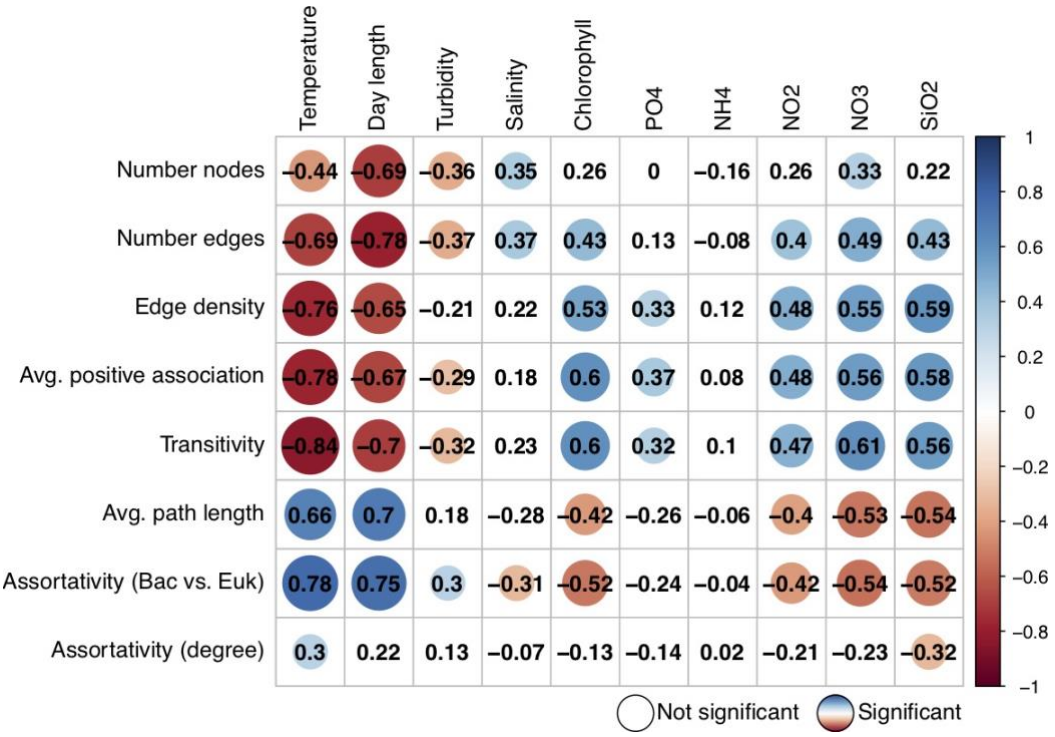


Figure 9**: Correlation Analysis through linear regression**. Using the temporal network, we correlated six global network metrics with environmental factors including the nutrients PO_4^{3-}, NH_4^+, NO_2^-, NO_3^- and SiO_2. The global network metrics were: Edge density, Average positive association (Avg. pos. ass.) score, Transitivity, Average path length (Avg. path length), Assortativity (degree), and Assortativity (bacteria vs. eukaryote). The number, circle's size and color in the square correspond to the Spearman correlation scores, no circle indicates non-significance.

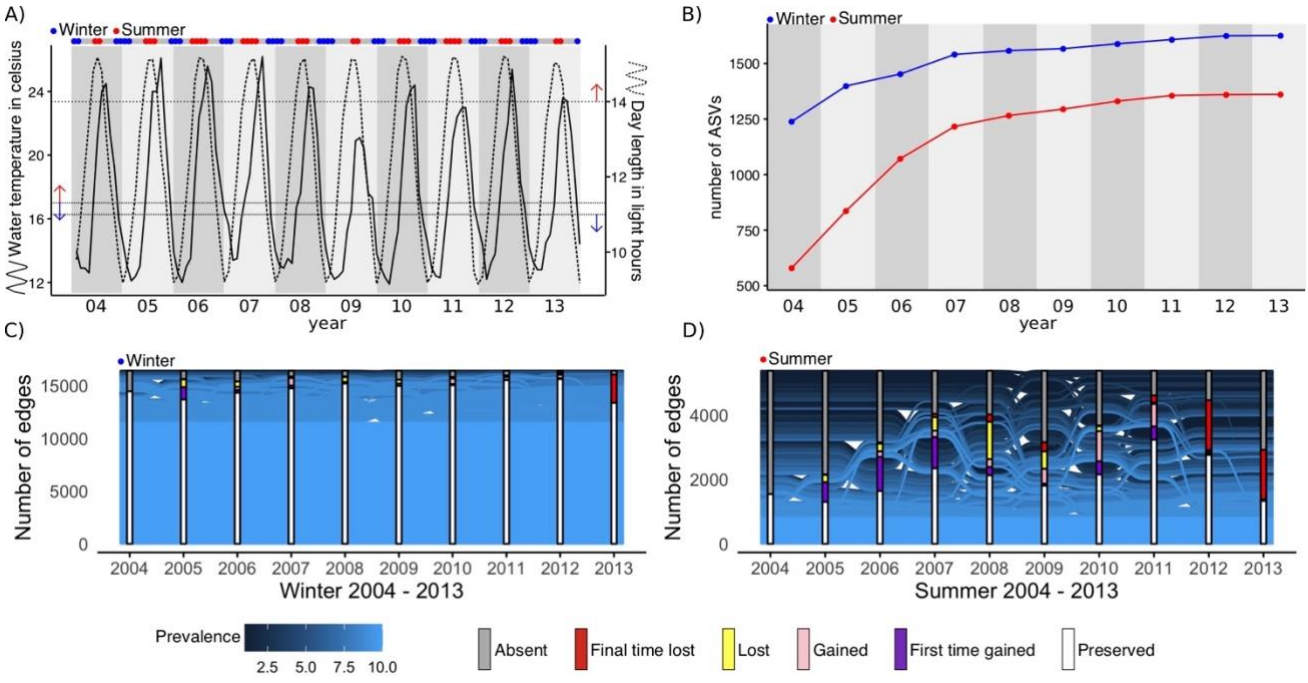


Figure 10**: Number of preserved, gained, and lost edges in summer and winter**. A) Indicates how we determined summer indicated with red dots (temperature above 17 °C and day length above 14 hours) and winter indicated with blue dots (temperature below 17 °C and day length below 11 hours); gray dots indicate months that are neither summer nor winter. B) accumulation curve of ASVs per year for winter (blue) and summer (red). C) and D) number of preserved, gained, and lost edges for winter and summer, respectively. The colors of flows indicate the prevalence of an edge with 10 (light blue) being present in each year, and 1 (dark blue) appearing in only one year. An edge appears in a year if it appears in at least one monthly subnetwork in the corresponding season. In winter, most edges appear in all years (light blue indicating 100% prevalence with edges present in all ten years), i.e. most edges are preserved in the consecutive months (we see a flow from the white preserved box to the next white preserved box). In summer, compared to winter, less edges are present in a month (combination of boxes indicating preserved, first time gained, and gained), and more edges are (re)gained and lost throughout the years (subsequently prevalence is lower indicated through darker blue).

Potential core associations

A single static network can comprise permanent, seasonal, and temporary associations. By comparing monthly subnetworks, we identified edges that remain (preserved), appear (gained), or disappear (lost) over time (Figure 7B). Intuitively, we would classify permanent associations through 100% recurrence. However, no association fulfilled the 100% criteria. Most associations had a low recurrence with three-quarters of the associations present in no more than 38% (46 monthly subnetworks). The average association prevalence increased slightly for taxonomically more related microorganisms (Figure 11). Considering the 100 most prevalent associations, which appeared in 71.7-98.3% (86-118) monthly subnetworks, 87 were bacterial associations (Table 9).

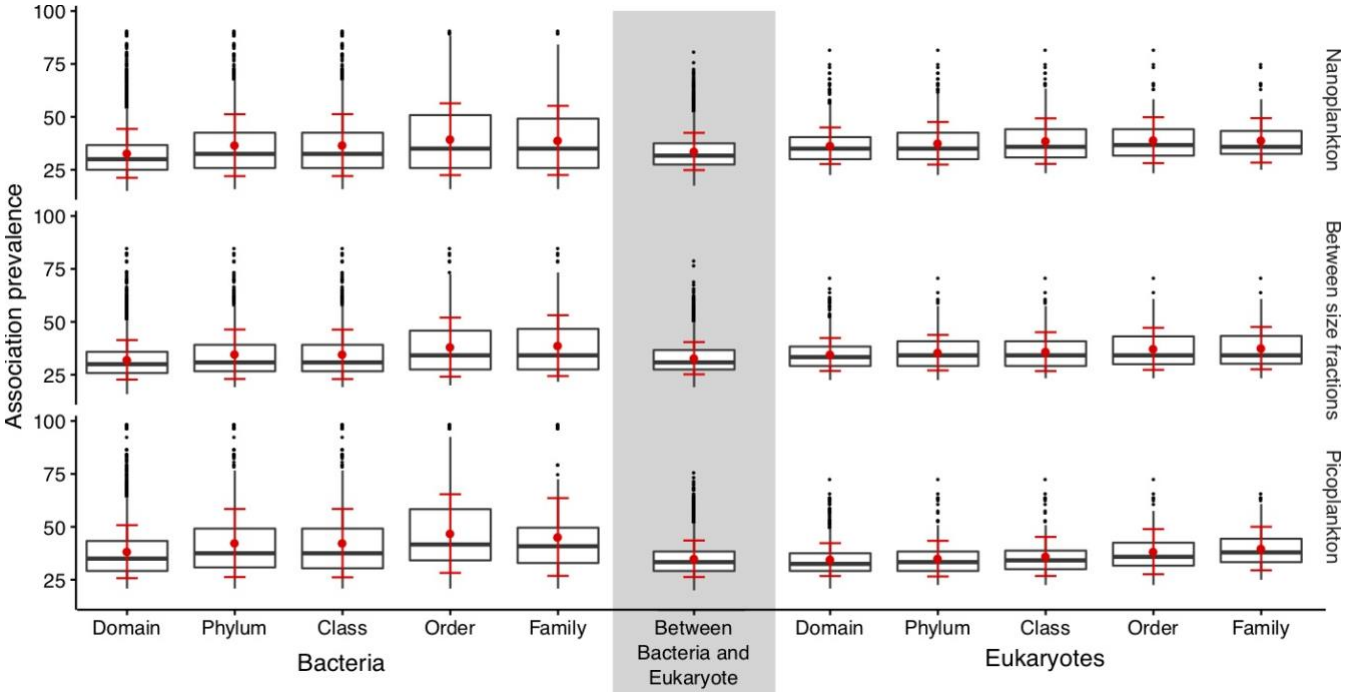


Figure 11: **Association prevalence increases slightly when microorganisms are taxonomically more related**. We grouped the associations according to the taxonomic classification of association partners (columns) and size fractions (rows). For example, "Class" groups associations between bacteria and eukaryotes, respectively, which were assigned to the same class. The gray column groups associations between bacteria and eukaryotes. The boxplot shows the association prevalence over a decade, i.e. in how many monthly subnetworks an association appears (given as fraction from 0 to 100% = 120 networks).

Table 9: **Top 100 most prevalent/recurring associations.** Associations were classified based on the domain of association partners.

Association partners	Number of associations
Bacterial association in picoplankton	42
Bacterial association in nanoplankton	35
Bacterial associations between size fractions	10
Bacteria associated to Eukaryote in nanoplankton	4
Eukaryotic association in nanoplankton	3
Bacteria associated to Eukaryote in picoplankton	3
Bacteria in nanoplankton associated to Eukaryotic picoplankton	2
Eukaryotic association in picoplankton	1

Although temporal recurrence of associations over the ten years was low, we found high recurrence in corresponding months from different years. We quantified the fraction of subnetworks in which each association appeared (Figure 12). We observed the highest prevalence from December to March, and the lowest prevalence from June to August (Figure 12). For each month, we taxonomically characterized prevalent associations appearing in at least nine out of ten monthly subnetworks (Figure 13). We found more association partners in colder waters compared to warmer waters. *Alphaproteobacteria* associations dominated, especially in April and May. The *Alphaproteobacteria* ASVs having highly prevalent associations belonged to *Pelagibacter ubique* (SAR11 Clades Ia & II), *Rhodobacteraceae*, *Amylibacter*, *Puniceispirillales (SAR116)*, *Ascidiaceihabitans*, *Planktomarina*, *Parvibaculales (OCS116)*, and *Kiloniella*. Between April and May, we noticed a large increase in the fraction of associations including *Cyanobacteria* or *Bacteroidetes* as association partners. While *Cyanobacteria* associations were a small fraction during November-April, they had a dominant role from May-October along with *Bacteroidetes* and *Alphaproteobacteria* associations (Figure 13).

Dynamic associations within main taxonomic groups: the case of Cyanobacteria

Our results indicated that associations are dynamic within specific taxonomic groups. Therefore, we investigated their behavior in *Cyanobacteria* given the importance of this group as primary producers in the ocean. We found 661 associations for *Cyanobium*, *Prochlorococcus*, and *Synechococcus* ASVs (Figure 14 and Figure 15). Most associations between cyanobacterial ASVs were positive (63 of 65), only a *Synechococcus* (referred to as bn_ASV_5) was negatively associated (association score measured -0.5) to other *Synechococcus* (bn_ASV_1 and bn_ASV_25), which were positively associated (association score of 0.8). While bn_ASV_5 appeared mainly in colder months, the other two appeared mainly in warmer months (Figure 15). All *Cyanobacteria* had more associations to other bacteria (in total 433) than eukaryotes (in total 163), which were dinoflagellate (103), Chlorophyta (25), Ochrophyta (12), Cryptophyta (11), Stramenopiles (5), Ciliophora (5), and Cercozoa (2).

Within the temporal network, the fraction of *Cyanobacteria* associations was highest in April-October (Figure 14A), which are the months with the fewest edges in the entire temporal network (Figure 7B), e.g., in the year 2011 (Figure 14B). We found that cyanobacterial ASVs, although being evolutionarily related, behaved differently in terms of number of associations over time, and association partners (Figure 15). For example, *Synechococcus* bn_ASV_5 had less partner than bn_ASV_1 according to numbers of associations but more according to taxonomic variety; both belonged to the most abundant ASVs (Figure 15). Only a tiny fraction of *Prochlorococcus* (e.g. bp_ASV_18) association partners were other *Cyanobacteria*, which contrasted to *Synechococcus* and *Cyanobium* (Figure 15). Moreover, we observed that *Cyanobium* (bn_ASV_20) connected to one *Deltaproteobacteria* (SAR324) ASV during the first eight years, but the association disappeared in the last two years. In particular, the inferred association duration was 101 months, starting March 2004 and ending with July 2012. After summer 2012, the Deltaproteobacteria ASV was not detected except from a few reads in November and December of 2012 and 2013. This *Cyanobacteria* example is likely representative of the dynamics of associations within other main taxonomic groups.

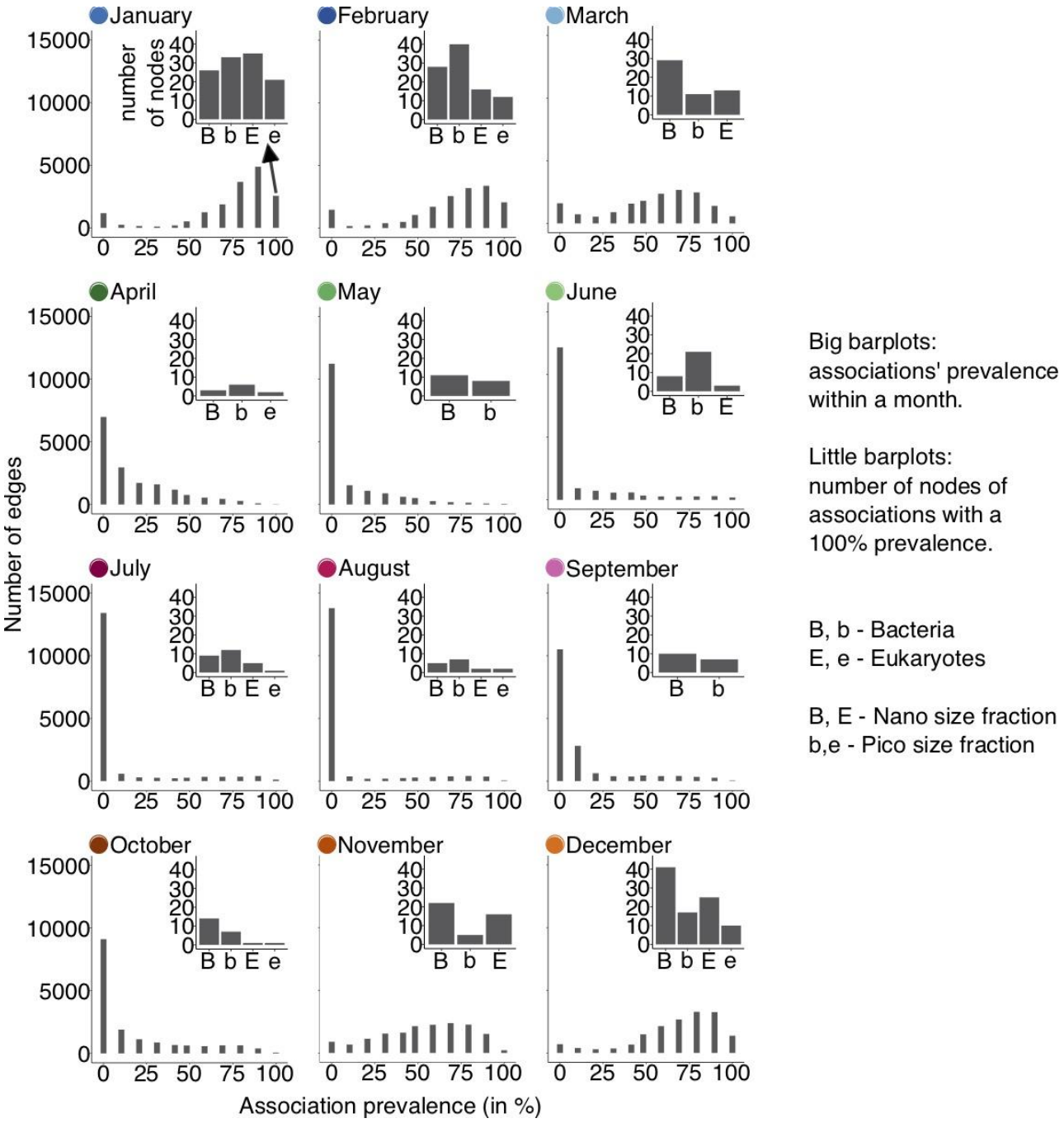


Figure 12**: Association prevalence per month.** Big bar plots: distribution of associations' prevalence for each month. For example, the bar at 100 for January indicates the number of edges that have been present in all Januarys of the ten-year time series. Small bar plots: number of nodes forming the associations with a 100% prevalence. For example, only bacteria were responsible for the edges during May, with an association prevalence of 100%. Bacteria are indicated with B or b, eukaryote with E or e. ASVs from the nano size-fraction have a capital letter (B, E), and ASVs from the pico size-fraction have a small letter (b, e).

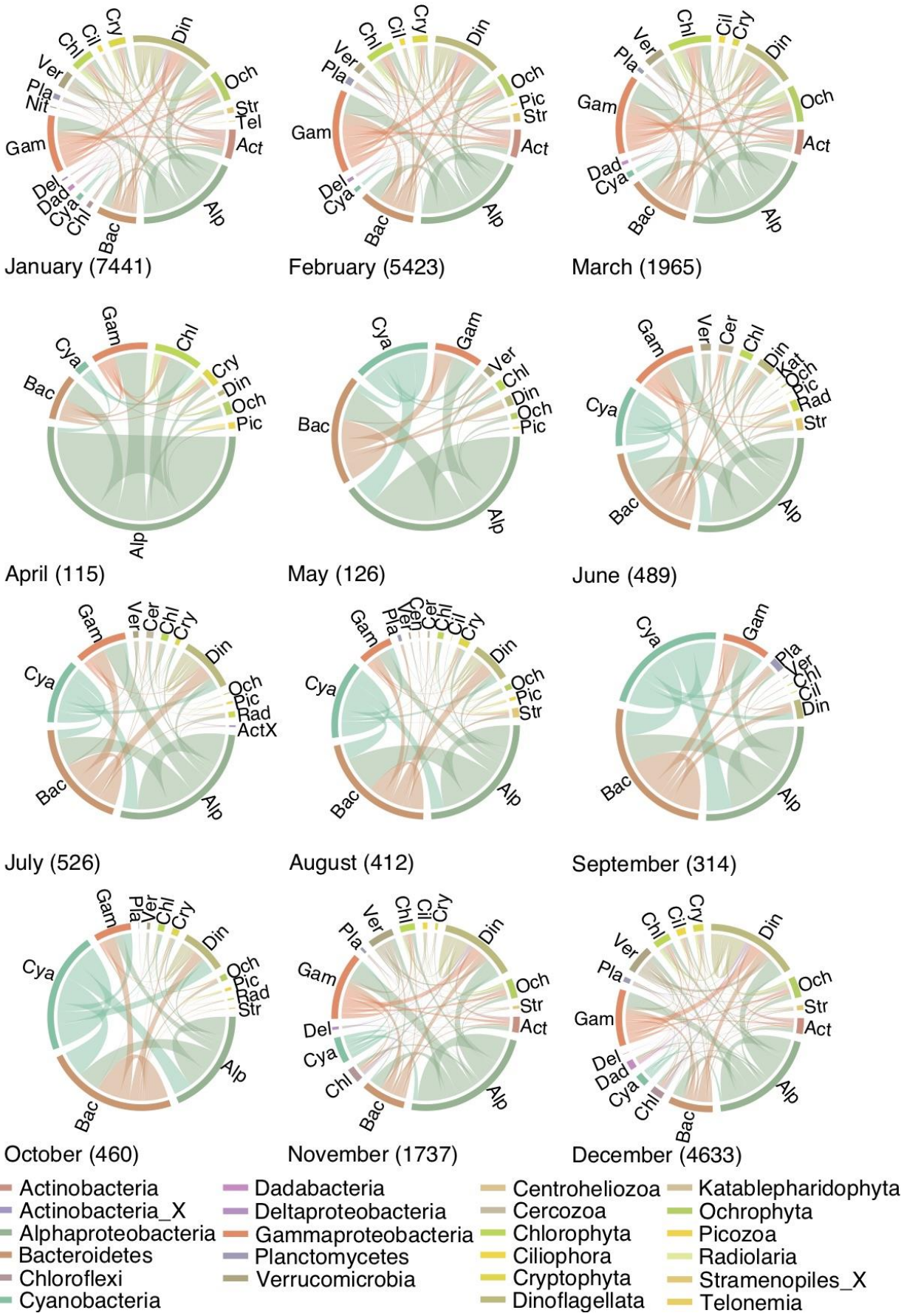


Figure 13: **Highly prevalent associations.** Associations with a monthly prevalence of at least 90%. Bacteria and eukaryotes are separated and ordered alphabetically. We provide in parentheses the number of associations that appeared in at least nine out of ten monthly subnetworks.

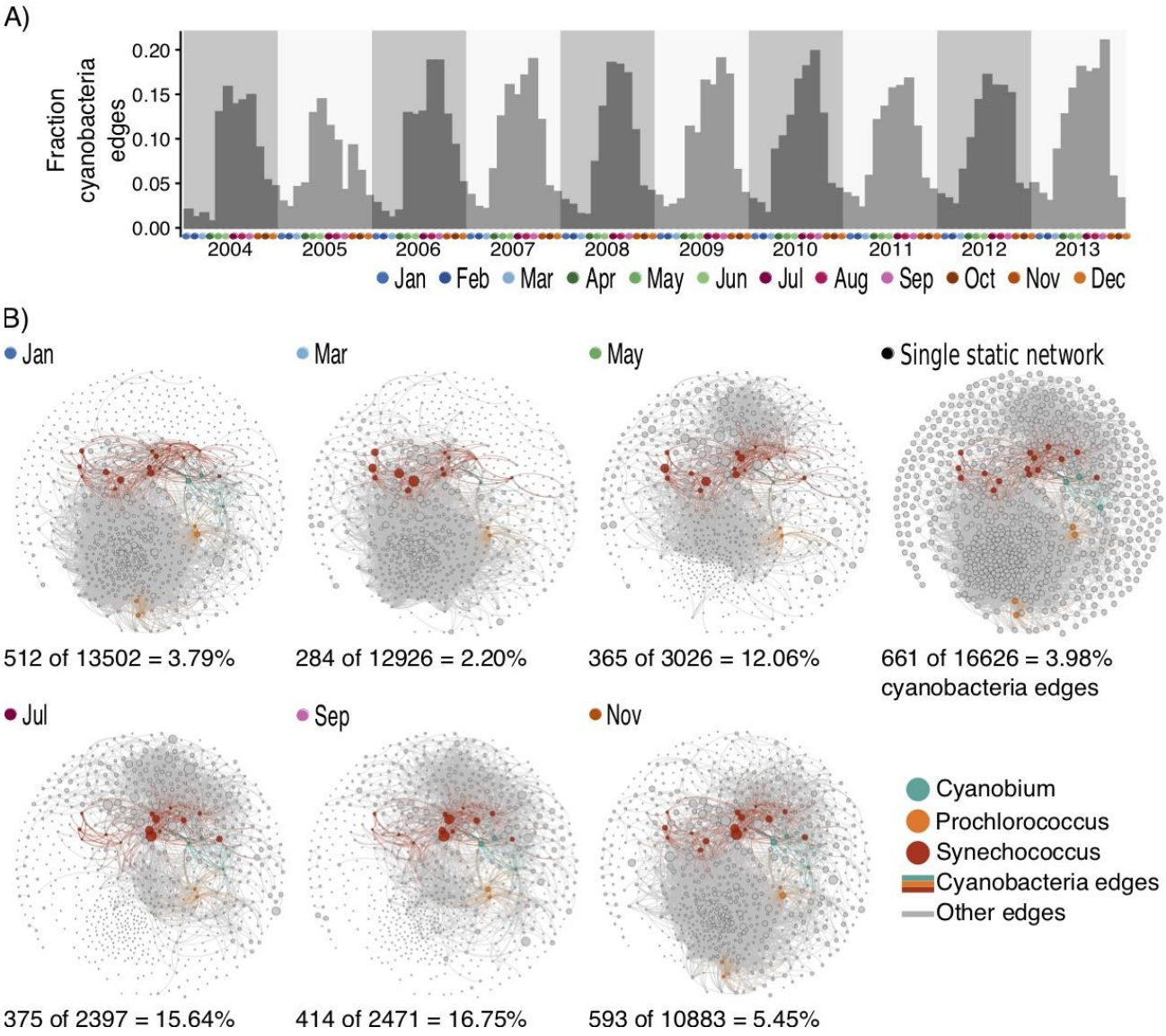
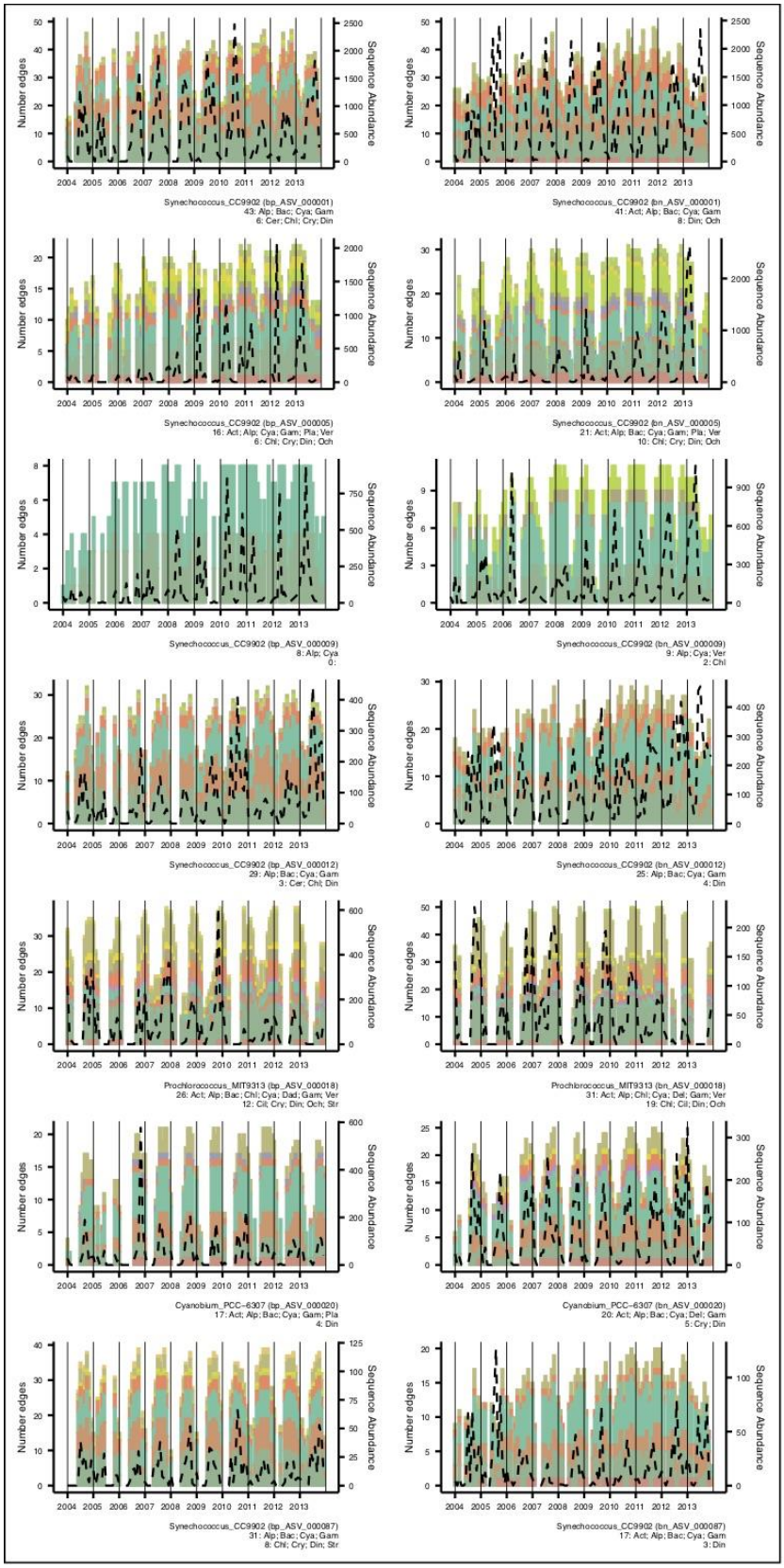


Figure 14: ***Cyanobacteria* associations**. A) Fraction of edges in the temporal network containing at least one *Cyanobacteria*. B) Location of *Cyanobacteria* associations in the temporal network and the single static network. Here we show, as an example, selected months of year 2011. The number and fraction of cyanobacterial edges and total number of edges is listed below each monthly subnetwork and the single static network.

(see next page)

Figure 15: **Association Partners of Cyanobacteria**. Number of Cyanobacteria associations in the temporal network (stacked bars) and the cyanobacterial sequence abundance in each month (black dashed line). Within the box, figures are split by ASVs (rows) and size fraction: picoplankton (left column) and nanoplankton (right column). The unboxed plots on the right are ASVs detected only in the nanoplankton. The height of the bar indicates the number of edges in each month for each cyanobacterial ASV. The color indicates the taxonomy of the association partner. From bottom to top, first appear bacteria and then eukaryotes, both sorted alphabetically. The subtitle shows the number of association partners followed by their identifiers (first 3 letters) for bacteria and eukaryotes.


Number edges
Sequence Abundance
2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Synechococcus_CC9902 (bp_ASV_000001)
43: Alp; Bac; Cya; Gam
6: Cer; Chl; Cry; Din
Synechococcus_CC9902 (bn_ASV_000001)
41: Act; Alp; Bac; Cya; Gam
8: Din; Och
Synechococcus_CC9902 (bn_ASV_000025)
38: Alp; Bac; Cya; Gam
12: Chl; Din; Str
Synechococcus_CC9902 (bp_ASV_000005)
16: Act; Alp; Cya; Gam; Pla; Ver
6: Chl; Cry; Din; Och
Synechococcus_CC9902 (bn_ASV_000005)
21: Act; Alp; Bac; Cya; Gam; Pla; Ver
10: Chl; Cry; Din; Och
Synechococcus_CC9902 (bn_ASV_000031)
8: Alp; Bac; Cya; Fir
2: Din
Synechococcus_CC9902 (bp_ASV_000009)
8: Alp; Cya
0:
Synechococcus_CC9902 (bn_ASV_000009)
9: Alp; Cya; Ver
2: Chl
Synechococcus_CC9902 (bn_ASV_000060)
11: Act; Bac; Cya; Gam; Pla; Ver
8: Chl; Din; Och
Synechococcus_CC9902 (bp_ASV_000012)
29: Alp; Bac; Cya; Gam
3: Cer; Chl; Din
Synechococcus_CC9902 (bn_ASV_000012)
25: Alp; Bac; Cya; Gam
4: Din
Synechococcus_CC9902 (bn_ASV_000079)
32: Act; Alp; Bac; Cya; Gam; Pla
15: Chl; Din; Str
Prochlorococcus_MIT9313 (bp_ASV_000018)
26: Act; Alp; Bac; Chl; Cya; Dad; Gam; Ver
12: Cil; Cry; Din; Och; Str
Prochlorococcus_MIT9313 (bn_ASV_000018)
31: Act; Alp; Chl; Cya; Del; Gam; Ver
19: Chl; Cil; Din; Och
Synechococcus_CC9902 (bn_ASV_000098)
8: Bac; Cya; Gam; Ver
3: Chl; Din
Cyanobium_PCC-6307 (bp_ASV_000020)
17: Act; Alp; Bac; Cya; Gam; Pla
4: Din
Cyanobium_PCC-6307 (bn_ASV_000020)
20: Act; Alp; Bac; Cya; Del; Gam
5: Cry; Din
Cyanobium_PCC-6307 (bn_ASV_000141)
16: Act; Alp; Bac; Cya; Del; Gam
8: Chl; Din
Synechococcus_CC9902 (bp_ASV_000087)
31: Alp; Bac; Cya; Gam
8: Chl; Cry; Din; Str
Synechococcus_CC9902 (bn_ASV_000087)
17: Act; Alp; Bac; Cya; Gam
3: Din
Prochlorococcus_MIT9313 (bn_ASV_000176)
75: Act; Alp; Bac; Chl; Cya; Dad; Del; Gam; Pla; Ver
20: Cil; Cry; Din; Och
Prochlorococcus_MIT9313 (bn_ASV_000267)
41: Act; Alp; Bac; Chl; Cya; Dad; Del; Gam; Pla; Ver
5: Cil; Din
Bacteria
Actinobacteria
Alphaproteobacteria
Bacteroidetes
Chloroflexi
Cyanobacteria
Dadabacteria
Deltaproteobacteria
Gammaproteobacteria
Planctomycetes
Verrucomicrobia
Eukaryote
Cercozoa
Chlorophyta
Ciliophora
Cryptophyta
Dinoflagellata
Ochrophyta
Stramenopiles_X

Discussion

Previous work found yearly recurrence in microbial community composition at the BBMO (Giner *et al.*, 2019; Auladell *et al.*, 2020; Krabberød *et al.*, 2021), and at the Bay of Banyuls (Lambert *et al.*, 2019), both in the NW Mediterranean Sea. Our approach focused in the connectivity of microorganisms and how they organize themselves from a network perspective. Similar to previous studies (Giner *et al.*, 2019; Lambert *et al.*, 2019; Auladell *et al.*, 2020; Krabberød *et al.*, 2021), our temporal network displayed seasonality with annual periodicity for most global network metrics. In general, our measured global network metrics are within previous work range (Steele *et al.*, 2011; Chow *et al.*, 2013, 2014; Cram, Xia, *et al.*, 2015; Lima-Mendez *et al.*, 2015; Zhao, Shen, *et al.*, 2016; Chaffron *et al.*, 2020) (Table 10 for edge density, transitivity, and average path length). Contrary to early works reporting biological networks generally being disassortative (negative assortativity based on degree) (Newman, 2002), our single static network and monthly subnetworks were assortative. Microorganisms had more and stronger connections and a tighter clustering in colder than in warmer waters. Seasonal bacterial freshwater networks (Zhao, Shen, *et al.*, 2016) also showed higher clustering in fall and winter than spring and summer, but in contrast to our work, networks were biggest in summer and smallest in winter. In agreement with our results, Chaffron *et al.* (2020) reported higher association strength, edge density, and transitivity in polar regions (colder) compared to other regions (warmer) of the global ocean. Colder waters in the Mediterranean Sea are milder than polar waters, but together, these results suggest that either microorganisms interact more in colder environments, or that their recurrence is higher due to higher environmental selection exerted by low temperatures and therefore, they tend to co-occur. Alternatively, lack of resources (mostly nutrients) in summer or in the tropical and subtropical ocean may prevent the establishment of several microbial interactions. In any case, temperature may not be the only driver of network architecture.

The effects of environmental variables on network metrics are unclear (Röttjers & Faust, 2018), yet, our approach allowed identifying potential environmental drivers of network architecture. Correlation analyses pointed to the usual suspects that have been already found to influence microbial abundances. For instance, our results indicated that temperature and day length, key variables driving microbial assemblages in seasonal time-series (Bunse & Pinhassi, 2017; Giner *et al.*, 2019; Lambert *et al.*, 2019), and to a lesser extent inorganic nutrients, were the main factors influencing global network metrics. This is also in agreement with earlier works indicating that phosphorus and nitrogen are the primary limiting nutrients in the Western Mediterranean Sea (Estrada, 1996; Sala *et al.*, 2002). Altogether, our correlation analysis is a step forward to elucidate the effects of environmental variables on network metrics, although we did not consider several other variables that could affect networks (e.g. organic matter).

Table 10: **Global network metrics of previously described microbial association networks.**

Edge density	Transitivity	Average path length	Sampling	Location	Domains	Notes	Reference
0.04	0.26	3.05	Monthly samples August 2000 - March 2004	Subsurface deep chlorophyll maximum depth off the southern California coast (SPOT)	Archaea, bacteria, and eukaryotes	Edge density for microbial network including environmental factors. Transitivity and average path length for microbial network.	(Steele et al., 2011)
0.14	0.33	1.94	Monthly samples August 2000 - January 2011	Two depths at SPOT	Free-living bacteria and some picoeukaryotes	Metrics from surface layer network.	(Chow et al., 2013)
0.02	0.24		Monthly samples March 2008 - January 2011	surface ocean (0-5m) at SPOT	Free-living eukaryotes (0.7–20 µm), bacteria (0.22–1 µm) and viruses (30 kDa–0.22 µm)		(Chow et al., 2014)
0.04	0.28	2.07	Monthly samples August 2003 - January 2011	Five depths at SPOT	Free-living bacteria	Metrics from 5 m layer network.	(Cram, Xia, et al., 2015)
(0.023) W:0.033 Sp:0.032 S:0.036 F:0.029	(0.472) W:0.518 Sp:0.480 S:0.475 F:0.573	(4.84) W:2.16 Sp:5.03 S:7.26 F:3.04	Spatial samples	52 samples from freshwater lakes (surface) in China	Bacteria	Metrics for (whole network) and seasonal networks: W: winter, Sp: spring, S: summer, and F: fall	(Zhao, Shen, et al., 2016)
E:0.005 EP:0.003 P:0.008	E:0.2 EP:0.0 P:0.43	E:3.05 EP:3.02 P:2.56	Spatial sampling	68 stations from the Tara Oceans expeditions (TARA) at two depths across eight oceanic provinces	Organisms from seven size fractions spanning from viruses to small metazoans	Metrics from surface networks including E: eukaryotes only, EP: eukaryotes and prokaryotes (0.5-5 µm), and P: prokaryotes only (0.2-1.6 µm)	(Lima-Mendez et al., 2015)
0.002	0.036		Spatial sampling	Samples from 115 stations from the TARA at two depths covering all major oceanic provinces from pole to pole	Bacteria, archaea, and eukaryotes from six size fractions.	Metrics represent the means of sample-specific subnetworks.	(Chaffron et al., 2020)

Our preliminary network (significant associations derived with eLSA) contained 18% negative edges compared to 0.9% in the single static network (after applying EnDED and Jaccard index). Thus, our filtering strategy removed proportionally more negative edges. Associations may represent positive or negative interactions, but they can also indicate high niche overlap (positive association) or divergent niches (negative association) between microorganisms (Hernandez *et al.*, 2021). We hypothesize that most of the removed negative edges represented associations between microorganisms from divergent niches, most likely corresponding to colder or warmer months.

We found more highly prevalent associations within specific months, than when considering all ten-years of data. Furthermore, our results indicate a potentially low number of core interactions and a vast number of non-core ones. Usually, core microorganisms are defined based on sequence abundances, as microorganisms (or taxonomical groups) appearing in all samples or habitats being under investigation (Shade & Handelsman, 2012). Shade & Handelsman (Shade & Handelsman, 2012) suggested other parameters, including connectivity, will create a more complex portrait of the core microbiome and advance our understanding of the role of key microorganisms and functions within and across ecosystems (Shade & Handelsman, 2012). Using a temporal network, we identified core associations based on recurrence, which contributes to our understanding of key interactions underpinning microbial ecosystem function. Considering associations within each month, we found more highly-prevalent associations in colder than in warmer months. Our results indicated microbial connectivity is more repeatable (indicating higher predictability) in colder than in warmer waters. On one hand, the microbial community in colder waters being more recurrent (Giner *et al.*, 2019) may explain our observations indicating a more robust connectivity. On the other hand, it may be the stronger connectivity that leads to more similar communities in colder waters in BBMO. Last but not least, the interplay of both species dynamics and interactions may determine community turnover in the studied ecosystem. From a technical viewpoint, the overall single static network may have missed to capture summer associations resulting in smaller monthly subnetworks. For instance, a previous work in freshwater lakes constructed season specific networks and found more associations in summer than winter with *Cyanobacteria* dominating in summer, which may be due to strong co-occurrence patterns and suitable living conditions (Zhao, Shen, *et al.*, 2016).

Several network-based analyses have been used to study *Cyanobacteria* associations. For example, Chow et al. (Chow *et al.*, 2014) determined for 12 *Cyanobacteria* (*Prochlorococcus* and *Synechoccus*) 44 potential relationships with two potential eukaryote grazers (a ciliate and a dinoflagellate), 39 to other bacteria and three between *Cyanobacteria*, which were all positive. Similarly, all cyanobacterial ASVs in our study connected primarily to other bacterial ASVs, and exerted mainly positive associations. In agreement, Cyanobacteria also displayed primarily positive associations in a network determined for the global ocean (Lima-Mendez *et al.*, 2015).

Identifying different potential association partners of closely related *Cyanobacteria*, may indicate adaptations to different niches. A recent study found distinct seasonal patterns of closely related taxa indicating niche partitioning at the BBMO, including *Synechococcus* ASVs (Auladell *et al.*, 2020). Our approach can complement and further characterize "sub"-niches by providing association partners for different ASVs.

Moreover, in contrast to a single static network, temporal networks allow identifying associated partners in time (Figure 15). An increase in abundance of a microorganism may promote the growth of associated partners and a decrease may hinder the growth of partners or cause predators to prey on other microorganisms. Moreover, given the majority of association partners being other bacteria, the growth of *Cyanobacteria* may affect other bacteria and their growth, which is why it is necessary to explore potential interaction partners (Zhao, Shen, *et al.*, 2016).

From a technical perspective, our approach allowed us to see what the single static network captured since all our temporal network observations are linked to it. Thus, future studies with higher sampling frequency may be able to construct networks within a month. However, our approach is a good starting point that allows us to move forward, but still, it has limitations, suggesting caution when making biological interpretations from the temporal network. Another limitation is that we disregarded local network patterns by using global network metrics. Future work could use the local-topological metric based on graphlets (Pržulj *et al.*, 2004). Counting the number of graphlets a node is part of quantifies their local connection patterns, which allows to infer seasonal microorganisms through recurring connection patterns in a temporal network. Such a network-based approach would complement the detection of the seasonal microorganisms based on sequence abundances (Giner *et al.*, 2019).

Conclusion

Incorporating the temporal dimension in the microbial association analysis unveiled multiple patterns that often remain hidden when using static networks. We developed a post-network-construction approach to generate a temporal network from a single static network that represents a step forward for disentangling the temporal nature of microbial associations. Yet, this approach has limitations, such as the monthly sampling frequency in our study. Using a higher sampling frequency would be the main solution. Investigating a coastal marine microbial ecosystem over ten years revealed a one-year-periodicity in the network topology. The temporal architecture was not stochastic, but displayed a modest amount of recurrence over time, especially in winter. Altogether, our approach allows comparing (sub)networks across spatiotemporal scales. Future efforts to understand the ocean microbiome should consider the dynamics of microbial interactions as these can be basis of ecosystem function.

Methods

The Blanes Bay Microbial Observatory (BBMO)

BBMO is a coastal oligotrophic site in the North-Western Mediterranean Sea (41°40′N 2°48′E) with not many identified natural disturbances and little anthropogenic pressures, with the exception of the construction of a nearby harbor from 2010 to 2012 (Gasol *et al.*, 2016; Ferrera *et al.*, 2020). The seasonal cycle is typical for a temperate coastal system (Gasol *et al.*, 2016), and the main environmental factors influencing microbial seasonal succession in temperate waters have been well studied and are known (Bunse & Pinhassi, 2017).

Shortly, the water column is slightly stratified in summer before it destabilizes and mixes in late fall, increasing the availability of inorganic nutrients with maximum concentrations in winter, between November and March. The high amounts of nutrients and increasing light induce phytoplankton blooms, mostly in late winter-early spring. During summer, inorganic nutrients become limiting, primary production is minimal, and dissolved organic carbon accumulates (Gasol *et al.*, 2016).

From sampling to sequence abundances

We sampled surface water (≈1m depth) monthly from January 2004 to December 2013 to determine microbial community composition and also measured ten environmental variables, which were previously described (Gasol *et al.*, 2016; Giner *et al.*, 2019): water temperature (°C) and salinity (obtained in situ with a SAIV-AS-SD204 Conductivity-Temperature-Depth probe), day-length (hours of light), turbidity (Secchi depth in meters), total chlorophyll-a concentration (μg/l, fluorometry of acetone extracts after 150 ml filtration on GF/F filters), and five inorganic nutrients: PO_4^{3-}, NH_4^+, NO_2^-, NO_3^- and SiO_2 (μM, determined with an Alliance Evolution II autoanalyzer (Grasshoff *et al.*, 2009)).

Sampling of microbial communities, DNA extraction, rRNA-gene amplification, sequencing and bioinformatic analyses are explained in detail in (Krabberød *et al.*, 2021). In short, 6 L of water were prefiltered through a 200 µm nylon mesh and subsequently filtered through another 20 µm nylon mesh and separated into nanoplankton (3 – 20 μm) and picoplankton (0.2 – 3 μm) using a 3 µm and 0.2 µm pore-size polycarbonate and Sterivex filters, respectively. Then, the DNA was extracted from the filters using a phenol-chloroform protocol (Schauer *et al.*, 2003), which has been modified for using Amicon units (Millipore) for purification. We amplified the 18S rRNA genes (V4 region) with the primers TAReukFWD1 and TAReukREV3 (Stoeck *et al.*, 2010), and the 16S rRNA genes (V4 region) with Bakt 341F (Herlemann *et al.*, 2011) and 806RB (Apprill *et al.*, 2015). Amplicons were sequenced in a MiSeq platform (2x250bp) at RTL Genomics (Lubbock, Texas). Read quality control, trimming, and inference of Amplicon Sequence Variants (ASVs) was made with DADA2 (Callahan *et al.*, 2016), v1.10.1, with the maximum number of expected errors set to 2 and 4 for the forward and reverse reads, respectively.

Microbial sequence abundance tables were obtained for each size fraction for both microbial eukaryotes and prokaryotes. Before merging the tables, we subsampled each table to the lowest sequencing depth of 4907 reads with the *rrarefy* function from the Vegan R-package (Oksanen *et al.*, 2019), v2.4-2, (see details in (Krabberød *et al.*, 2021)). We excluded 29 nanoplankton samples (March 2004, February 2005, May 2010 - July 2012) due to suboptimal amplicon sequencing. In these, abundances were estimated using seasonally aware missing value imputation by the weighted moving average for time series as implemented in the imputeTS R-package (Moritz & Gatscha, 2017), v2.8.

Sequence taxonomy was inferred using the naïveBayesian classifier method (Wang *et al.*, 2007) together with the SILVA database (Quast *et al.*, 2012), v.132, as implemented in DADA2 (Callahan *et al.*, 2016). Additionally, eukaryotic microorganisms were BLASTed (Altschul *et al.*, 1990) against the Protist Ribosomal

Reference (PR2) database (Guillou *et al.*, 2012), v4.10.0. The PR2 classification was used when the taxonomic assignment from SILVA and PR2 disagreed. We removed ASVs that identified as Metazoa, Streptophyta, plastids, mitochondria, and Archaea since the 341F-primer was not optimal for recovering this domain (McNichol *et al.*, 2020).

The resulting table contained 2924 ASVs, Table 11A. Next, we removed rare ASVs, keeping ASVs with sequence abundance sums above 100 reads and prevalence above 15% of the samples, i.e., we considered taxa present in at least 19 months. The resulting table contained 1782 ASVs, Table 11B. An ASV can appear twice, in the nano and pico size fractions due to dislodging cells or particles and filter clogging. This can introduce biases in our analysis. To reduce these biases, as done previously (Krabberød *et al.*, 2021), we divided the abundance sum of the bigger by the smaller size-fraction 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. This operation removed two eukaryotic ASVs and 41 bacterial ASVs from the nanoplankton, and 30 bacterial ASVs from the picoplankton (Table 11C). The resulting table was used for network inference.

Table 11**: Number and fraction of ASVs and reads. We list the number of ASVs, and the** total, bacterial and eukaryotic number of reads for the sequence abundance tables before removing rare ASVs (A), after removing rare ASVs (B), and after the size-fraction filtering (C), the preliminary network with significant edges (D), and the single static network obtained after removing environmentally-driven edges and edges with association partners appearing more often alone than with the partner (E). If an ASV appeared in the nano- and pico-plankton size fractions, it was counted twice.

Count tables	ASVs	Reads	Eukaryote	Eukaryotic reads	Bacteria	Bacterial reads
A	2 924	2 273 548	1 365	1 121 855	1 559	1 151 693
B	1 782	2 155 318	1 009	1 057 599	773	1 097 719
C	1 709	2 062 866	1 007	1 057 263	702	1 005 603
D	754	1 657 885	306	730 025	448	927 860
E	709	1 621 959	294	719 558	415	902 401
Fractions	ASV	Reads	Eukaryote	Eukaryotic reads	Bacteria	Bacterial reads
B/A*100	60.94	94.80	73.92	94.27	49.58	95.31
C/A*100	58.45	90.73	73.77	94.24	45.03	87.32
D/C*100	44.12	80.37	30.39	69.05	69.05	92.27
E/C*100	41.49	78.63	29.20	68.06	59.12	89.74

A – raw sequence abundance table; B – sequence abundance table without rare ASVs; C –sequence abundance table after size-fraction filtering; D – preliminary network with significant edges; E – single static network

From sequence abundances to the single static network

First, we constructed a preliminary network using the tool eLSA (Xia *et al.*, 2011, 2013), as done in (Deutschmann *et al.*, 2020; Krabberød *et al.*, 2021), including default normalization and z-score transformation, using median and median absolute deviation. Although we are aware of time-delayed interactions, we considered

our 1-month sampling interval as too large for inferring time-delayed associations with a solid ecological basis, and focused on contemporary interactions between co-occurring microorganisms. Using 2000 iterations, we estimated p-values with a mixed approach that performs a random permutation test of a co-occurrence if the comparison's theoretical *p*-values are below 0.05. The Bonferroni false discovery rate (*q*) was calculated based on the *p*-values using the *p.adjust* function from the stats R-package (R Core Team, 2019). We used the 0.001 significance threshold for the *p* and *q* values, as suggested in other studies (Weiss *et al.*, 2016). We refrained from using an association strength threshold since it may not be appropriate to differentiate between true interactions and environmentally-driven associations (Deutschmann *et al.*, 2020), and changing thresholds have been shown to lead to different network properties (Connor *et al.*, 2017). The preliminary network contained 754 nodes and 29820 edges (24458, 82% positive, and 5362, 18% negative).

Second, for environmentally-driven edge detection, we applied EnDED (Deutschmann *et al.*, 2020), combining the methods Interaction Information (with a 0.05 significance threshold and 10000 iterations) and Data Processing Inequality. We inserted artificial edges connecting each node to each environmental parameter. We identified and removed 3315 (11.12%) edges that were environmentally-driven, i.e., 26505 edges (23405, 88.3% positive, and 3100, 11.7% negative) remained (Table 12 and Table 13).

Table 12: **Number of environmental factors leading to the removal of edges**.

Number of environmental factors	Edges	Positive edges	Negative edges
no environmentally-driven edges	26505	23405 (88.3%)	3100 (11.7%)
1	2747	1019 (37.1%)	1728 (62.9%)
2	506	33 (6.5%)	473 (93.5%)
3	61	1 (1.6%)	60 (98.4%)
4	1	0 (0%)	1 (100%)

Table 13: **Environmentally-driven edges for each environmental factor**. Number of environmentally-driven edges and their fraction considering the total number of edges (29820) in the network. In addition, we present the number of positive and negative edges and their fraction considering number of edges removed through an environmental factor.

Environmental factor	Edges	Positive edges	Negative edges
Temperature	1920 (6.44%)	725 (37.8%)	1195 (62.2%)
Total chlorophyll-a concentration	838 (2.81%)	82 (9.8%)	756 (90.2%)
Day length	730 (2.45%)	237 (32.5%)	493 (67.5%)
NO_2^-	192 (0.64%)	26 (13.5%)	166 (86.5%)
SiO_2	162 (0.54%)	6 (3.7%)	156 (96.3%)
NO_3^-	57 (0.19%)	12 (21.1%)	45 (78.9%)
Turbidity	47 (0.16%)	0	47 (100%)
Salinity, NH_4^+, and PO_4^{3-}	0	0	0

Third, we determined the Jaccard index, J, for each microorganism pair associated through an edge. Let S_i be the set of samples in which both microorganisms are present (sequence abundance above zero), and S_u be the set of samples in which one or both microorganisms are present. Then, we can calculate the Jaccard index as the fraction of samples in which both appear (intersection) from the number of samples in which at least one appears (union): $J = S_i/S_u$. We chose $J > 0.5$, which removed 9879 edges and kept 16626 edges (16481, 99.1%

positive and 145, 0.9% negative). We removed isolated nodes, i.e., nodes without an associated partner in the network. The number and fraction of retained reads are listed in Table 11. The resulting network is our single static network.

From the single static network to the temporal network

We determined the temporal network comprising 120 sample-specific (monthly) subnetworks through the three conditions indicated below and visualized in Figure 6. The subnetworks are derived from the single static network and contain a node subset and an edge subset of the static network. Let e be an association between microorganisms A and B, with association duration $d = (t_1, t_2)$, i.e., the association starts at time point t_1 and ends at t_2. Then, considering month m, the association e is present in the monthly subnetwork N_m, if

1) e is an association in the single static network,
2) the microorganisms A and B are present within month m, and
3) m is within the duration of association, i.e., $t_1 \leq m \leq t_2$.

With the 2nd condition, we assumed that an association was present in a month if both microorganisms were present, i.e., the microbial abundances were non-zero for that month. However, we cannot assume that microbial co-occurrence is a sufficient condition for a microbial interaction because different mechanisms influence species and interactions, and the environmental filtering of species and interactions can be different (Poisot *et al.*, 2012). Using only the species occurrence assumption would increase association prevalence. To lower this bias, we also required that the association was present in the static network, 1st condition, and within the association duration, 3rd condition, both inferred by eLSA (Xia *et al.*, 2011, 2013). Lastly, we removed isolated nodes from each monthly subnetwork.

Network analysis

We computed global network metrics to characterize the single static network and each monthly subnetwork, using the igraph R-package (Csardi & Nepusz, 2006). Some metrics tend to be more correlated than others implying redundancy between them and clustering them into four groups (Jamakovic & Uhlig, 2008). Thus, we selected one metric from each group: *edge density*, *average path length*, *transitivity*, and *assortativity* based on node degree. In addition, we also computed the *average strength of positive associations* between microorganisms using the mean, and *assortativity* based on the nominal classification of nodes into bacteria and eukaryotes. Assortativity (bacteria vs. eukaryotes) is positive if bacteria tend to connect with bacteria and eukaryotes tend to connect with eukaryotes. It is negative if bacteria tend to connect to eukaryotes and vice versa. We also quantified associations by calculating their prevalence as the fraction of monthly subnetworks in which the association was present for all ten years (recurrence), and monthly. We visualized highly prevalent associations with the circlize R-package (Gu *et al.*, 2014). We tested our hypotheses of environmental factors influencing network topology by calculating the Spearman correlations between global network metrics and environmental data, using Holm's multiple test correction to adjust p-values (Holm, 1979), with the function

corr.test in psych R-package (Revelle, 2020). We used Gephi (Bastian *et al.*, 2009), v.0.9.2, and the Fruchterman Reingold Layout (Fruchterman & Reingold, 1991) for network visualization.

Cyanobacteria

Our dataset contained 19 cyanobacterial ASVs, which all appeared in the nano-, and nine in the picoplankton. This is against expectations, as *Cyanobacteria* are part of the pico-plankton. Yet, they have been observed in fractions above 3 µm at BBMO (Mestre *et al.*, 2020). Recovering ASVs in the nanoplankton may be due to cell aggregation, particle attachment, clogging of filters or being prey to bigger microorganisms. We blasted the sequences against the Cyanorak database (Garczarek *et al.*, 2021), v.2. We performed BLASTN matches against the nucleotide database containing all *Synechococcus* and *Prochlorococcus* RNAs with the -evalue 1.0e-5 option. We found 2812 sequences comprising 95 different ecotypes (considering name, clade and subclade), with 93.84-100% identity. There were 63 sequences (34 different microorganisms) with a similarity of 100% for 11 ASVs. Most matching sequences were found for *Synechococcus* ASV_1. While *Synechococcus* ASV_5 had only two 100% hits, they did not 100% match to ASV_1 (Table 14).

Table 14: **Cyanobacterial ASVs**. 100% Matching sequences from Cyanorak database for selected cyanobacterial ASVs

ASV	Number	Matching sequence name with clade and subclade
Synechococcus #1	38	2x A15-24 III IIIa, 2x A15-28 III IIIb, 3x A15-44 II IIa, 2x A15-62 II IIc, 2x A18-40 III IIIa, 2x A18-46.1 III IIIa, 2x BOUM118 III IIIa, 2x CC9605 II IIc, 2x M16.1 II IIa, 2x PROS-U-1 II IIh, 2x ROS8604 I Ib, 3x RS9902 II IIa, 3x RS9907 II IIa, 2x RS9915 III IIIa, 2x TAK9802 II IIa, 1x WH8016 I Ib, 2x WH8103 III IIIa, 2x WH8109 II IIa
Synechococcus #5	2	2x PROS-9-1 I Ib
Prochlorococcus #18	2	1x EQPAC1 HLI HLI, 1x MED4 HLI HLI
Cyanobium #20	2	1x MINOS11 5.3 5.3, 1x RCC307 5.3 5.3

Availability of data and material

The BBMO microbial sequence abundances (ASV tables), taxonomic classifications, environmental data including nutrients will be publicly available after acceptance. The data are of course available upon request. Networks are already available. R-Markdowns for data analysis including commands to run eLSA and EnDED (environmentally-driven-edge-detection and computing Jaccard index) are publicly available: https://github.com/InaMariaDeutschmann/TemporalNetworkBBMO.

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 (ESR2, http://www.singek.eu) to RL. 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), MicroEcoSystems (240904, RCN, Norway) and MINIME (PID2019-105775RB-I00, AEI, Spain) to RL. FL was supported by the Spanish National Program FPI 2016 (BES-2016-076317, MICINN, Spain). SC was supported by the CNRS MITI through the interdisciplinary program Modélisation du Vivant (GOBITMAP grant). SC and DE were

supported by the H2020 project AtlantECO (award number 862923). A range of projects from the EU and the Spanish Ministry of Science funded data collection and ancillary analyses at the BBMO.

Author's contributions

The overall project was conceived and designed by RL and AKK. VB, JMG, and RM were responsible for the sampling and contextual data at the BBMO. RL processed the amplicon data from BBMO generating ASV tables. AKK constructed the initial preliminary network. It was the starting point of the present study, which is part of the overall project. IMD developed the conceptual approach and DE, SC, and RL contributed to its finalization. IMD performed the data analysis. ED, DE, SC, LFB, FL, AKK, CM, and RL contributed with biological interpretation of the results. IMD wrote the original draft. All authors contributed to manuscript revisions and approved the final version of the manuscript.

Acknowledgements

We thank all members of the Blanes Bay Microbial Observatory team with the multiple projects funding this collaborative effort. Part of the analyses have been performed at the Marbits bioinformatics core at ICM-CSIC (https://marbits.icm.csic.es).

Final remarks

- ⇒ Associations could represent permanent, temporary, or seasonal associations.
- ⇒ Monthly sampling does not allow the construction of one network per month.
- ⇒ Our post-network-construction approach allows determining monthly subnetworks derived from an overall single static network.
- ⇒ Our approach can be used to quantify temporal recurrence for each association over the ten years and also for each month.
- ⇒ Most associations appear in colder months.
- ⇒ Associations are more repeatable at colder compared to warmer months.
- ⇒ The temporal network comprising the 120 monthly subnetworks appears to collapse from colder to warmer months and reassemble from warmer to colder months.
- ⇒ High prevalent associations may infer core associations that are essential for ecosystem functioning.
- ⇒ Our approach allows quantifying associations based on temporal recurrence, which was suggested in Chapter 5 to may strengthen and reduce the number of potential interaction hypotheses.
- ⇒ Proposed idea: our approach of quantifying edge recurrence (temporal) can be adjusted to determine spatial recurrence.
- ⇒ Note: in the previous chapter we used EnDED for edges that are within at least one (network-based) environmental triplet. Here, we inserted artificial edges (resulting in artificially-generated triplets) to use EnDED on each edge with each possible environmental factor. The number of identified indirect associations increased slightly.

⇒ Observation: we shortly scratched upon the application of using the temporal network to determine association partners in time for an example group (Cyanobacteria). In the future, such approach may be valuable for microbial ecologist studying specific organismal groups. However, this was outside the scope of the subproject and it was not the aim of this thesis, which focuses on network improvement and identifying general patterns.

Chapter 7 Disentangling marine microbial networks across space

Ina Maria Deutschmann, Erwan Delage, Caterina R. Giner, Marta Sebastián, Julie Poulain, Javier Arístegui, Carlos M. Duarte, Silvia G. Acinas, Ramon Massana, Josep M. Gasol, Damien Eveillard, Samuel Chaffron and Ramiro Logares

Abstract

Although microbial interactions underpin ocean ecosystem functions, they remain barely known. Different studies have analyzed microbial interactions using static association networks based on omics-data. However, microbial associations are dynamic and can change across physicochemical gradients and spatial scales, which needs to be considered to understand the ocean ecosystem better. We explored associations between archaea, bacteria, and picoeukaryotes along the water column from the surface to the deep ocean across the northern subtropical to the southern temperate ocean and the Mediterranean Sea by defining sample-specific subnetworks. Quantifying spatial association recurrence, we found the lowest fraction of global associations in the bathypelagic zone, while associations endemic of certain regions increased with depth. Overall, our results highlight the need to study the dynamic nature of plankton networks and our approach represents a step forward towards a better comprehension of the biogeography of microbial interactions across ocean regions and depth layers.

Keywords: association network; sample-specific subnetworks; microbial interactions; biogeography of associations; archaea, bacteria, and micro-eukaryotes; ocean

Introduction

Microorganisms play fundamental roles in ecosystem functioning (DeLong, 2009; Krabberød *et al.*, 2017) and ocean biogeochemical cycling (Falkowski *et al.*, 2008). The main processes shaping microbial community composition are selection, dispersal, and drift (Vellend, 2020). Selection exerted via environmental conditions and biotic interactions are essential in structuring the ocean microbiome (Logares *et al.*, 2020), leading to heterogeneities reflecting those in the ocean environment, mainly in terms of temperature, light, pressure, nutrients and salinity. In particular, global-scale studies of the surface ocean reported strong associations between microbial community composition and diversity with temperature (Sunagawa *et al.*, 2015; Ibarbalz *et al.*, 2019; Salazar *et al.*, 2019; Logares *et al.*, 2020). Marked changes in microbial communities with ocean depth have also been reported (Cram, Xia, *et al.*, 2015; Parada & Fuhrman, 2017; Mestre *et al.*, 2018; Peoples *et al.*, 2018; Xu *et al.*, 2018; Giner *et al.*, 2020), reflecting the steep vertical gradients in light, temperature, nutrients and pressure.

Prokaryotes (bacteria and archaea) and unicellular eukaryotes are fundamentally different in terms of ecological roles, functional versatility, and evolutionary history (Massana & Logares, 2013) and are connected through biogeochemical and food web interaction networks (Layeghifard *et al.*, 2017; Seymour *et al.*, 2017). Still, knowledge about these interactions remains limited despite their importance to understand better microbial

life in the oceans (Krabberød *et al.*, 2017; Bjorbækmo *et al.*, 2019). Such interactions are very difficult to resolve experimentally, mainly because most microorganisms are hard to cultivate (Baldauf, 2008; Lewis *et al.*, 2020) and synthetic laboratory communities are unlikely to mirror the complexity of wild communities. However, metabarcoding approaches to identify and quantify marine microbial taxa allow to infer microbial association networks, where nodes represent microorganisms and edges potential interactions.

Association networks provide a general overview of the microbial ecosystem aggregated over a given period of time (Steele *et al.*, 2011; Chow *et al.*, 2013, 2014; Cram, Xia, *et al.*, 2015; Needham *et al.*, 2017; Parada & Fuhrman, 2017) or through space (Lima-Mendez *et al.*, 2015; Milici *et al.*, 2016; Chaffron *et al.*, 2020). Previous work characterized potential marine microbial interactions, including associations within and across depths. For example, monthly sampling allowed investigating prokaryotic associations in the San Pedro Channel, off the coast of Los Angeles, California, covering the water column from the surface (5 m) to the seafloor (890 m) (Cram, Xia, *et al.*, 2015; Parada & Fuhrman, 2017). Furthermore, a global spatial survey occurring within the TARA Oceans expedition, allowed to investigate planktonic associations between a range of organismal size fractions in the epipelagic zone, from pole to pole (Lima-Mendez *et al.*, 2015; Chaffron *et al.*, 2020). However, these studies did not include the bathypelagic realm, below 1000 m depth, which represents the largest microbial habitat in the biosphere (Arístegui *et al.*, 2009).

A single static network determined from spatially distributed samples over the global ocean captures global, regional and local associations. Also, given that global-ocean expeditions collect samples over several months, networks could include temporal associations, yet, disentangling them from spatial associations is normally complicated and not considered. Global associations may constitute the core interactome, that is, the set of microbial interactions essential for the functioning of the ocean ecosystem (Shade & Handelsman, 2012). Core associations may be detected by constructing a single network from numerous locations and identifying the most significant associations and strongest associations (Coutinho *et al.*, 2015). On the other hand, regional and local associations may point to interactions occurring in specific spatial areas of different sizes due to particular taxa distributions resulting from environmental selection, dispersal limitation, ecological niches or biotic/abiotic filtering. The fraction of regional associations may be determined by excluding all samples belonging to one region and recomputing network inference with the reduced dataset (Lima-Mendez *et al.*, 2015). Alternatively, regional networks can be built allowing to determine both, global and regional associations (Mandakovic *et al.*, 2018) by investigating which edges networks have in common and which are unique. Such regional networks could contribute to understanding how the architecture of potential microbial interactions changes with environmental heterogeneity, also helping to comprehend associations that are stable (i.e., two partners always together) or variable (one partner able to interact with multiple partners across locations).

Regional networks, however, require a high number of samples per delineated zone, but these may not be available due to logistic or budgetary limitations. Recent approaches circumvent this limitation by deriving sample-specific subnetworks from a single static, i.e., all-sample network, which allows quantifying association recurrence over spatiotemporal scales (Chaffron *et al.*, 2020; Deutschmann *et al.*, 2021). Here, we adjusted this approach and used it to determine global and regional associations along vertical and horizontal ocean scales,

which allowed us determining the biogeography of marine microbial associations. We analyzed associations between archaea, bacteria, and picoeukaryotes covering the water column, from surface to deep waters, in the Mediterranean Sea (hereafter MS) and five ocean basins: North and South Atlantic Ocean, North and South Pacific Ocean, and Indian Ocean (hereafter NAO, SAO, NPO, SPO, and IO). We estimated microbial taxa abundances using 397 globally distributed samples from the epipelagic to the bathypelagic zone in six ocean regions (Figure 16). We separated most epipelagic samples into surface and deep-chlorophyll maximum (DCM) samples. Next, we constructed a first global network comprising 5457 nodes and 31966 edges, 30657 (95.9%) positive and 1309 (4.1%) negative. Then, we applied a filter strategy including the removal of environmentally-driven edges due to nutrients (4.9% NO_3^-, 4.2% PO_4^{3-}, 2.0% SiO_2), temperature (1.9%), salinity (0.2%), and Fluorescence (0.01%) (Table 15). Altogether, our sample-specific network-based exploration allowed us to determine core associations in the global ocean and specific regions, analyze changes in associations and network topology with depth and regions, and to investigate the vertical connectivity of connected planktonic associations.

Table 15**: Number of environmentally-driven edges detected by EnDED**. We removed environmentally-driven edges (indirect) from the preliminary network, which contained 31966 edges. Only edges that were not environmentally-driven by any environmental factor (not indirect) remained in the network.

Environmental factor	Number of samples	indirect	Not indirect
Fluorescence	394	4 (0.01%)	31962
NO3	361	1563 (4.9%)	30403
PO4	359	1357 (4.2%)	30609
Salinity	395	67 (0.2%)	31899
SiO4	360	632 (2.0%)	31334
Temperature	395	622 (1.9%)	31344
All		2848 (8.9%) = 1779 removed by 1 + 751 removed by 2 + 308 removed by 3 + 10 removed by 4	29118 (91.1%)

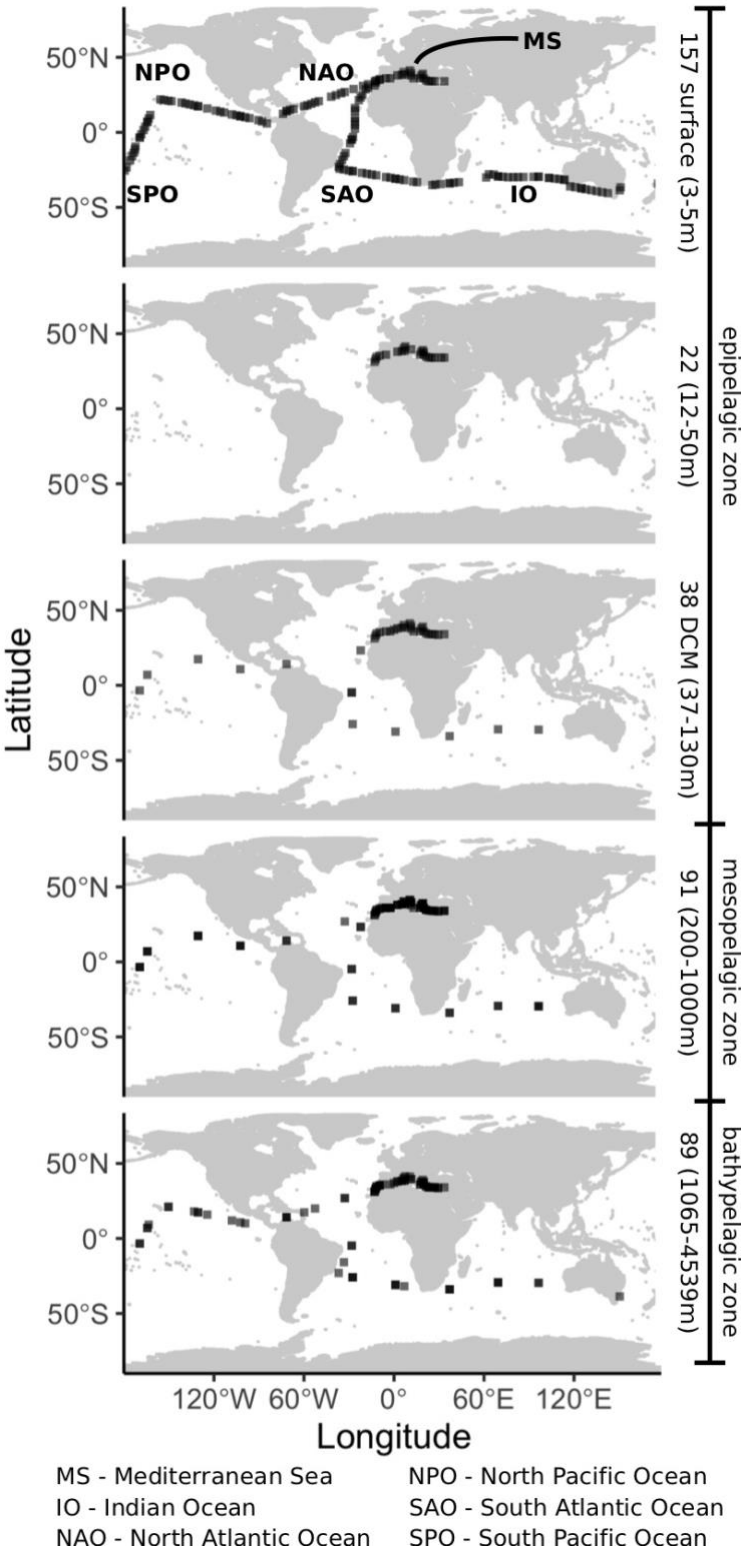


Figure 16**: Sampling scheme.** Location, number, and depth range of samples from the epipelagic zone including surface and DCM layer, the mesopelagic zone, and the bathypelagic zone from the global tropical and subtropical ocean and the Mediterranean Sea.

Results

From a global static network to sample-specific subnetworks

The resulting global static network contained 5448 nodes and 29118 edges, 28178 (96.8%) positive and 940 (3.2%) negative. It served as the underlying structure from which we generated 397 sample-specific subnetworks following three criteria. First, we required that an edge must be present in the global static network. Second, an edge can only be present within a subnetwork if both microorganisms associated with the edge have a sequence abundance above zero in the corresponding sample. Third, microorganisms associated need to appear together (intersection) in more than 20% of the samples, in which one or both appear (union) for that specific region and depth. This third condition was robust since random subsets retained most associations compared with the associations obtained when using all samples (Figure 17). In addition to these three conditions, a node is present in a subnetwork if it has at least one association partner. Consequently, each subnetwork is included in the global static network.

Spatial recurrence

We determined the spatial recurrence of each association using its prevalence computed as the fraction of subnetworks in which a given association was present across the 397 samples (Figure 18A) and within each region-depth-layer combination (Figure 18B). The global ocean surface layer (contributing with 40% of samples) had more associations compared to the other depths (Figure 18B). Remarkably, 14971 of 18234 (82.1%) global ocean surface associations were absent from the MS. In turn, the number of surface associations was similar across ocean basins (Figure 18B).

Considering the most prevalent associations (those found in over 70% of subnetworks), we found that major vertical taxonomic patterns were conserved across regions: the epipelagic layers (surface and DCM) and the two lower layers (meso- and bathypelagic zones) were more similar to each other, respectively (Figure 19). The fraction of associations including *Alphaproteobacteria* was moderate to high in all zones in contrast to *Cyanobacteria* appearing mainly, as expected, in the epipelagic zone (Figure 19). The fraction of *Dinoflagellata* associations was moderate to high in the epipelagic zone and lower in the meso- and bathypelagic zones. While *Dinoflagellata* associations dominated most epipelagic layers, fewer were found in the MS and SAO surface and NAO DCM (Figure 19). *Thaumarchaeota* associations were moderate to high especially in the mesopelagic (dominant in the MS), moderate in the bathypelagic, and lower in the epipelagic zone (Figure 19). Another interesting pattern is the increase in associations including *Gammaproteobacteria* with depth being higher in the meso- and bathypelagic than in the epipelagic, especially in the SAO, SPO, NPO and IO.

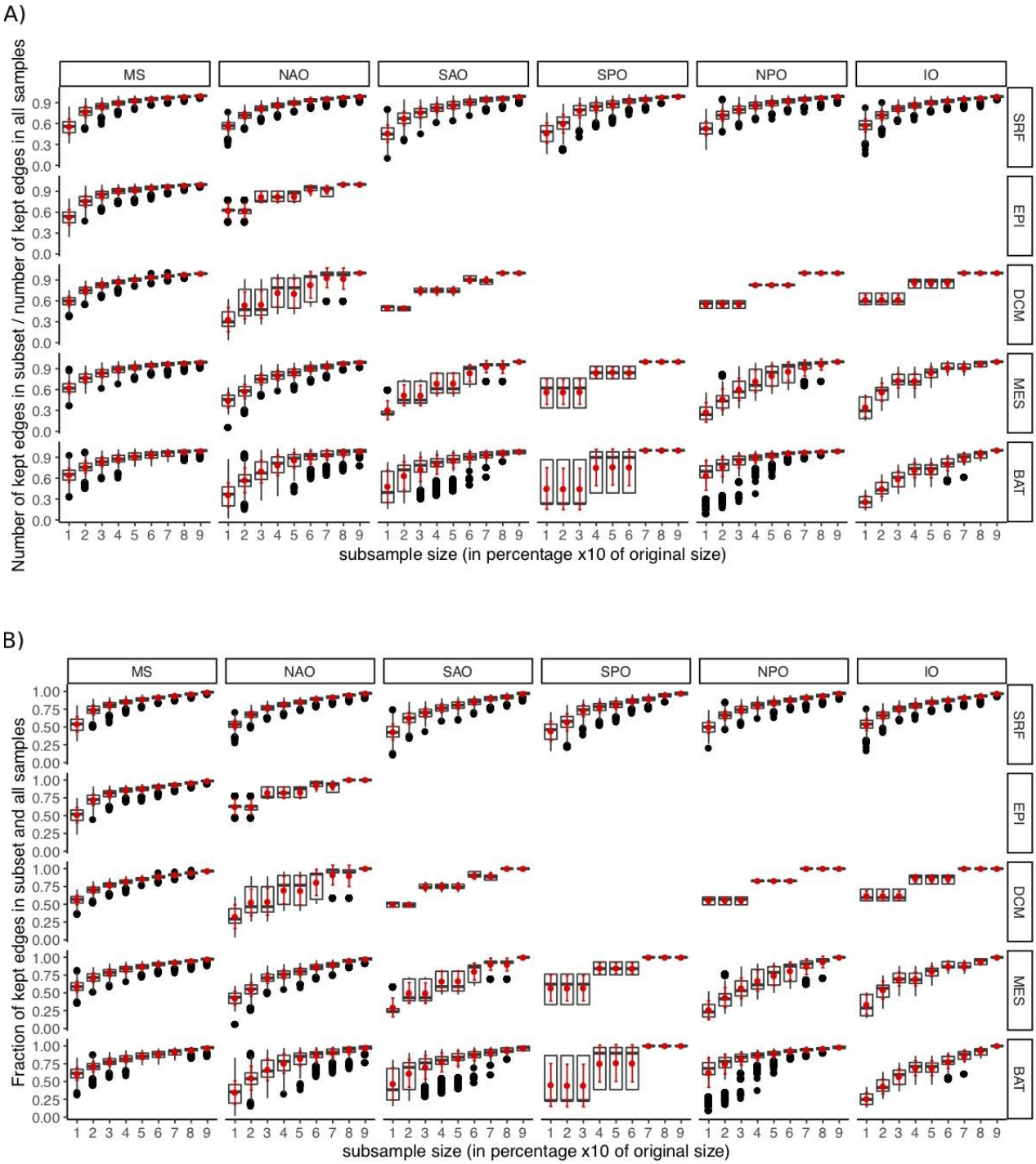
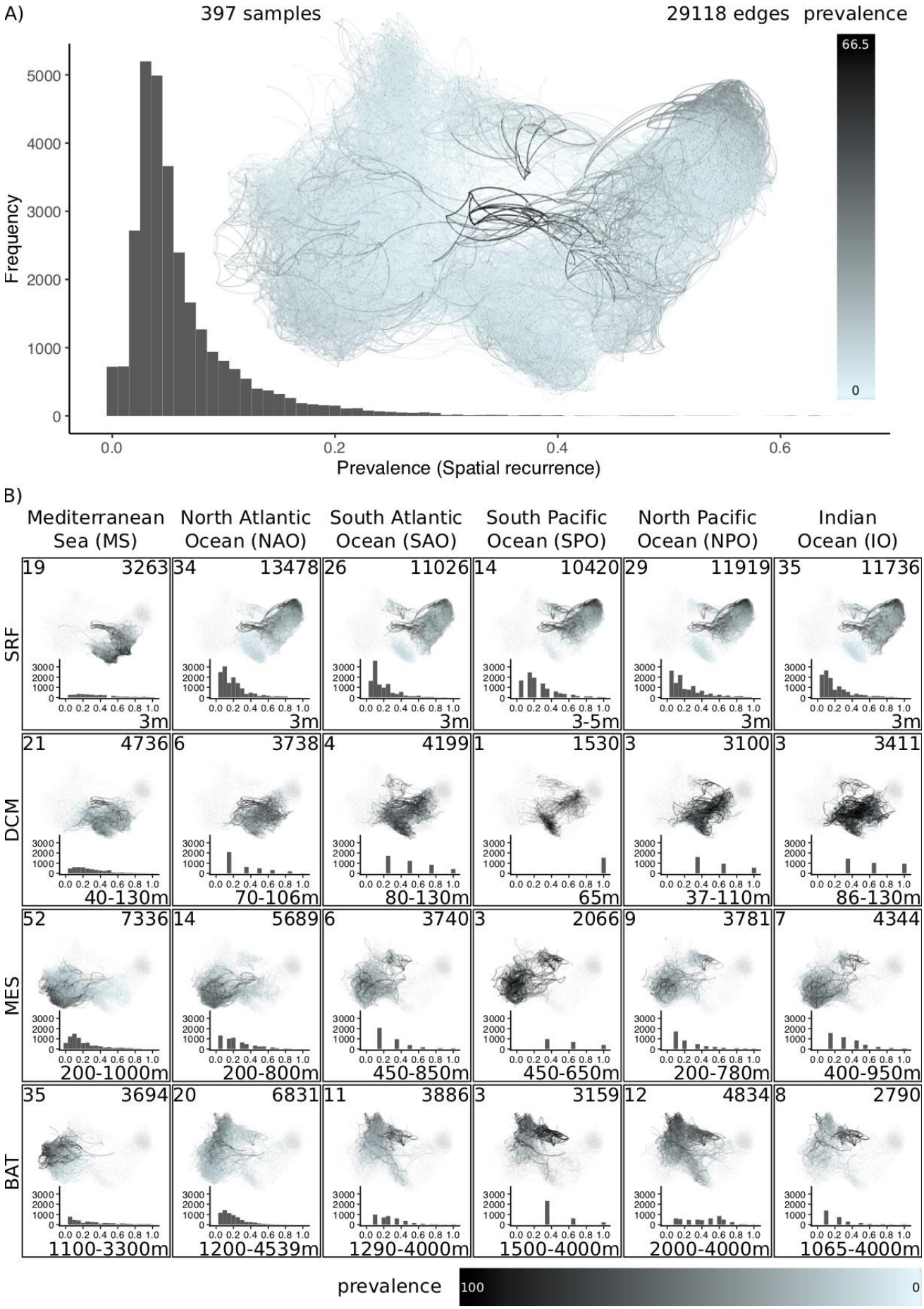


Figure 17**: Robustness of the third condition.** We tested the robustness of the third condition for generating sample-specific subnetworks for each region and depth with sufficient samples. The DCM layer from the SPO was removed because it contained only one sample. Within each region and depth, the set of samples was randomly subsampled containing between 10% to 90% of the samples in the original set using all samples. The y-axis shows the fraction of edges that were kept in the subsampled set compared to the original set. We considered A) only the number of kept edges and B) which edges were kept.

(see next page)

Figure 18**: Spatial recurrence**. A) Association prevalence showing the fraction of subnetworks in which an association appeared considering all depth layers across the global tropical and subtropical ocean and the Mediterranean Sea. Associations that occurred more often (black) appeared in the middle of the single static network visualization. Most edges had a low prevalence (blue) <20%. B) The sample-specific subnetworks of the four ocean layers (rows): surface (SRF), DCM, mesopelagic (MES), and bathypelagic (BAT), and the six regions (columns). The histograms show the association prevalence within each depth layer and region (excluding absent associations, i.e., 0% prevalence). The number of samples appears in the upper left corner, the number of edges with a prevalence >0% in the upper right corner, and the depth range in the lower right corner (in m below surface). Note that the prevalence goes up to 100% in B) vs. 66.5% in A).


A)
397 samples
29118 edges
prevalence
66.5
0
Frequency
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
Prevalence (Spatial recurrence)
B)
Mediterranean Sea (MS)
North Atlantic Ocean (NAO)
South Atlantic Ocean (SAO)
South Pacific Ocean (SPO)
North Pacific Ocean (NPO)
Indian Ocean (IO)
SRF
DCM
MES
BAT
19 3263 3m
34 13478 3m
26 11026 3m
14 10420 3-5m
29 11919 3m
35 11736 3m
21 4736 40-130m
6 3738 70-106m
4 4199 80-130m
1 1530 65m
3 3100 37-110m
3 3411 86-130m
52 7336 200-1000m
14 5689 200-800m
6 3740 450-850m
3 2066 450-650m
9 3781 200-780m
7 4344 400-950m
35 3694 1100-3300m
20 6831 1200-4539m
11 3886 1290-4000m
3 3159 1500-4000m
12 4834 2000-4000m
8 2790 1065-4000m
prevalence
100
0

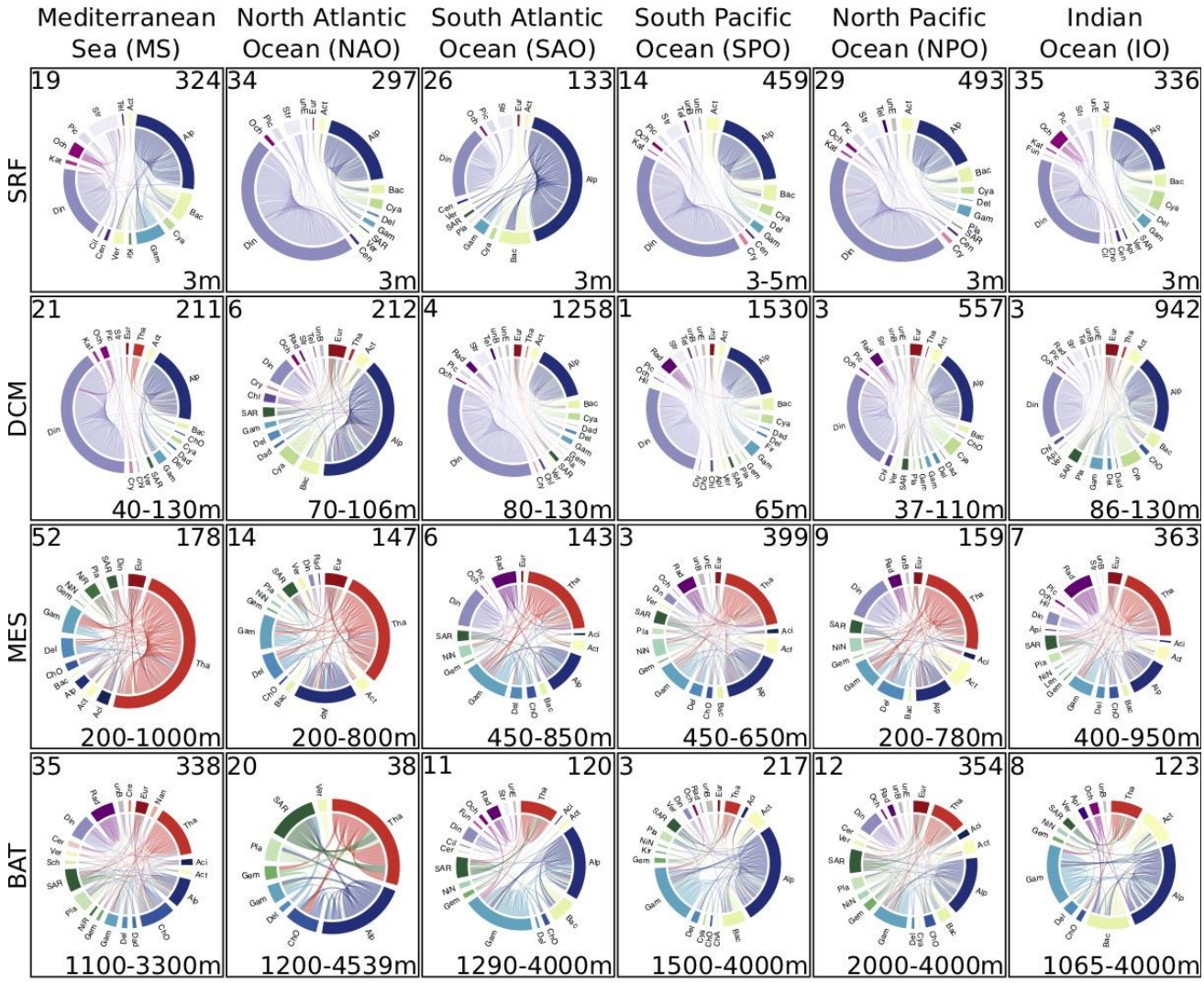


Figure 19**: Highly prevalent associations for each region and depth layer**. If an association appears in more than 70% of subnetworks it is classified as highly prevalent. The four ocean layers (rows) are surface (SRF), DCM, mesopelagic (MES), and bathypelagic (BAT). The number of samples appears in the upper left corner, the number of edges in the upper right corner, and the depth range in the lower right corner (in m below surface).

Highly prevalent associations present across all regions are candidates to represent putative core interactions in the global ocean, which are likely to perform processes crucial for ecosystem functioning. We defined global associations as those appearing in more than 70% of subnetworks in each region. While we found several (21-26) global associations in the epi- and mesopelagic zones, no global associations were identified in the bathypelagic zone (Table 16, Figure 20). In addition, we resolved prevalent (>50%) and low-frequency (>20%) associations. These three types of associations are distinct by definition, i.e., a global association cannot be assigned to another type. The fraction of global, prevalent, and low-frequency associations was highest in the DCM layer and lowest in the bathypelagic zone (third and fifth column in Table 16, Figure 20B, and Figure 20D). Given that the MS bathypelagic is warmer (median temperature of 13.78°C) than the global ocean bathypelagic (median temperature between 1.4°C in SPO and 4.41°C in NAO), we calculated these associations for the global ocean only. We found slightly to moderately more global, prevalent, and low-frequency associations in the global ocean when not considering the MS (fifth to seventh row in Table 16, Figure 20E-H).

Table 16**: Number of classified associations per depth layer.** The sum of classified associations (including Other) is the number of present associations. Absent associations appear in other layers but in no subnetwork of a given layer. Global, prevalent, and low-frequency associations have been computed with and without considering the MS. The proportion of regional associations increased with depth (row highlighted in gray).

Depth layer	Epipelagic (Surface)	Epipelagic (DCM)	Mesopelagic	Bathypelagic
Global	26 (0.14%)	23 (0.31%)	21 (0.20%)	-
Prevalent	22 (0.12%)	47 (0.64%)	10 (0.10%)	7 (0.07%)
Low-frequency	105 (0.58%)	160 (2.17%)	212 (2.05%)	51 (0.51%)
Global (no MS)	86 (0.47%)	52 (0.70%)	28 (0.27%)	9 (0.09%)
Prevalent (no MS)	207 (1.14%)	76 (1.03%)	27 (0.26%)	28 (0.28%)
Low-frequency (no MS)	1361 (7.46%)	219 (2.97%)	342 (3.30%)	489 (4.84%)
Regional	2014 (11.05%)	2290 (31.03%)	3420 (33.00%)	3669 (36.33%)
MS	596 (3.27%)	1295 (17.55%)	2254 (21.75%)	1217 (12.05%)
NAO	577 (3.16%)	306 (4.15%)	422 (4.07%)	1522 (15.07%)
SAO	162 (0.89%)	304 (4.12%)	301 (2.90%)	143 (1.42%)
SPO	152 (0.83%)	105 (1.42%)	40 (0.39%)	109 (1.08%)
NPO	298 (1.63%)	133 (1.80%)	204 (1.97%)	516 (5.11%)
IO	229 (1.26%)	147 (1.99%)	199 (1.92%)	162 (1.60%)
Other*	16067 (88.12%)	4860 (65.85%)	6701 (64.66%)	6372 (63.10%)
Other (no MS)*	14566 (79.88%)	4743 (64.27%)	6547 (62.17%)	55904 (58.46%)
Present	18234 (100%)	7380 (100%)	10364 (100%)	10099 (100%)
Absent	10884	21738	18754	19019

** The number of unclassified (Other) associations is computed from present, regional, global, prevalent, and low-frequency associations. The last three classifications have been done with and without the MS, and subsequently the number of unclassified (other) associations varies.*

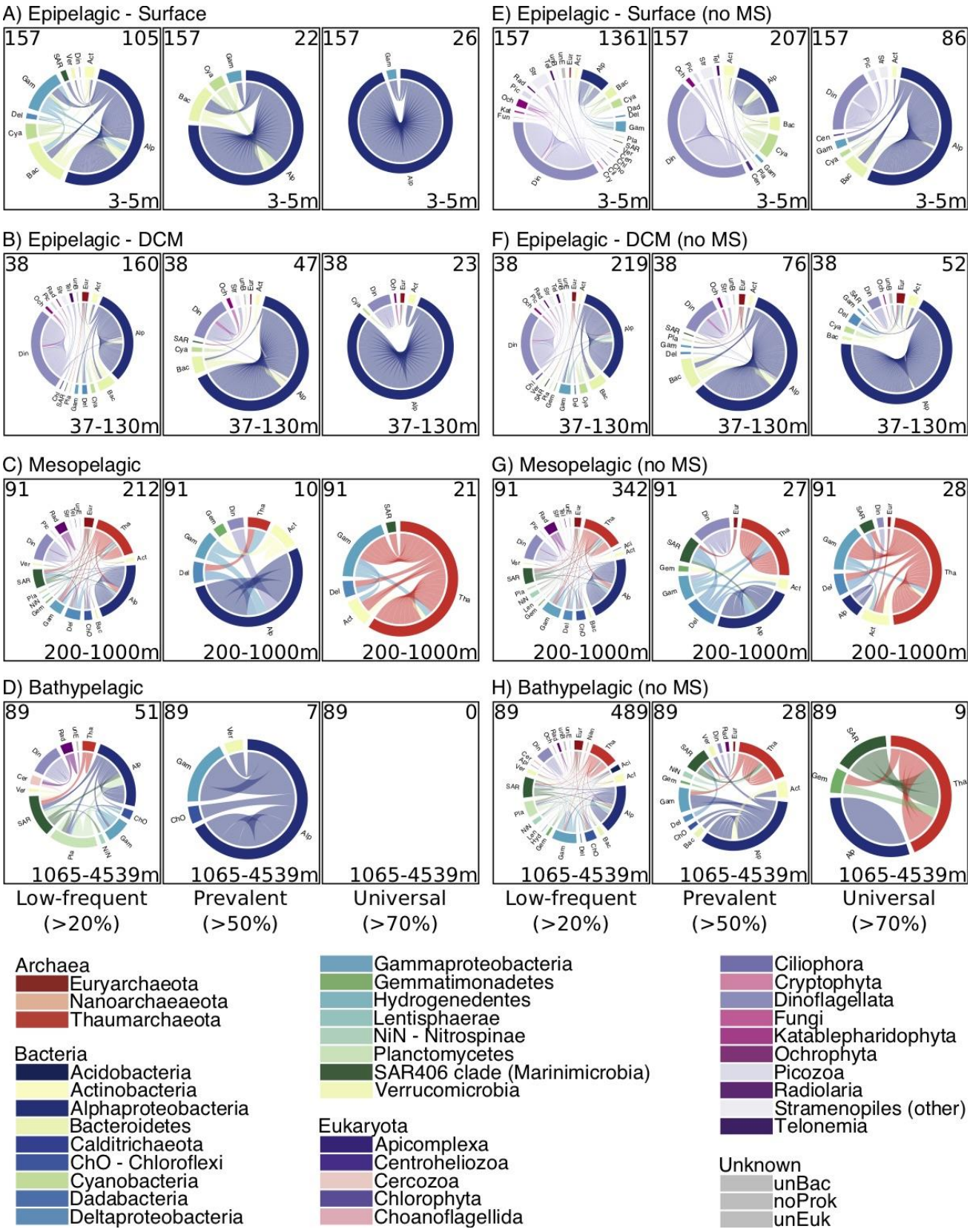


Figure 20**: Associations occurring in each region and depth layer**. If an association appears in more than 20% of subnetworks in each region, it is classified as low-frequency, >50% prevalent, and >70% global. The number of samples appears in the upper left corner, the number of edges in the upper right corner, and the depth range in the lower right corner (in m below surface). We classified the associations considering all six regions (A-D) and considering the five ocean basins not considering the the MS (E-H).

Next, we determined regional associations within each depth layer. A regional association was defined as detected in at least one sample-specific subnetwork of one region and absent from all subnetworks of the other five regions. Results indicated an increasing proportion of regional associations with depth (Table 16, Figure 21A-B, Figure 22). We found substantially more associations in the DCM and mesopelagic layers of the MS than corresponding layers of the global ocean. This may reflect the different characteristics of these layers in the MS vs. the global ocean or the massive differences in spatial dimensions between the global ocean and the MS. More surface and bathypelagic regional associations corresponded to the MS and NAO than in other regions (Table 16). Most regional associations had low prevalence, i.e., they were present in a few sample-specific subnetworks within the region (Figure 21C). We found 235 prokaryotic highly prevalent (>70%) regional associations in contrast to 89 eukaryotic and 24 associations between domains (Supplementary Material 1[8]).

Previous studies have found a substantial vertical connectivity in the ocean microbiota, with surface microorganisms having an impact in deep sea counterparts (Mestre et al., 2018; Ruiz-González et al., 2020). Thus, here, we analyzed the vertical connectivity of microbial associations. Few associations appeared throughout the water column within a region: 327 prokaryotic, 119 eukaryotic, and 13 associations between domains (Supplementary Material 2[9]). In general, most associations appearing in the meso- and bathypelagic did not appear in upper layers except for the MS and NAO where most and about half, respectively, of the bathypelagic associations already appeared in the mesopelagic (Figure 23, Table 17). Specifically, 81.77 – 90.90% mesopelagic and 43.54-72.71% bathypelagic associations appeared for the first time in the five ocean basins (Table 17). In the MS, 71.24% mesopelagic and 22.44% bathypelagic associations appeared for the first time and 69.71% of bathypelagic associations already appeared in the mesopelagic (Table 17). This points to specific microbial interactions occurring in the deep ocean that do not occur in upper layers. In addition, while most surface associations also appeared in the DCM in the MS, most surface associations disappeared with depth in the five oceans (Figure 23) suggesting that most surface ocean associations are not transferred to the deep sea, despite microbial sinking (Mestre et al., 2018). In fact, we observed that most deep ocean ASVs already appeared in the upper layers (Figure 24), in agreement with previous work that has shown that a large proportion of deep sea microbial taxa are also found in surface waters, and that their presence in the deep sea is related to sinking processes (Mestre et al., 2018).

[8] Supplementary Material 1: Description at the end of the methods section.
[9] Supplementary Material 2: Description at the end of the methods section.

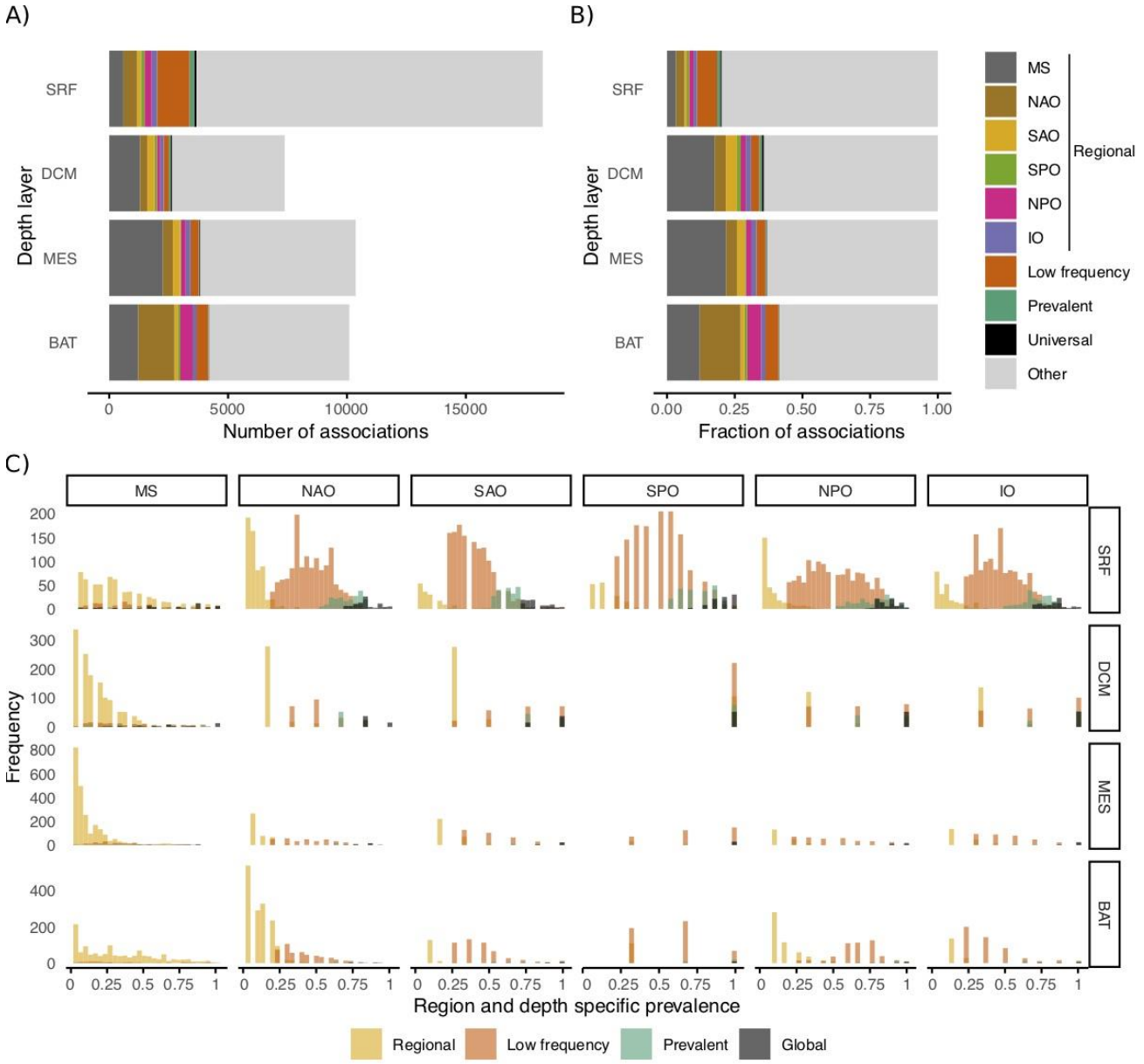


Figure 21**: Classification of associations**. An association can be classified into global (>70% prevalence, not considering the MS), prevalent (>50%, not considering the MS), low-frequency (>20%, not considering the MS), regional, and other. Regional associations are assigned to one of six ocean basins. The number A) and fraction B) of each type of association are shown for each depth layer: surface (SRF) and DCM (epipelagic), mesopelagic (MES) and bathypelagic (BAT). Color indicates the type of classification. The associations have been classified into the five types based on their prevalence in each region. The prevalence of associations is shown in C). For instance, global associations have a prevalence above 70% in each region (not considering the MS). Regional associations are present in one region (indicated with yellow with mainly low prevalence >0%) and absent in all other regions (0% prevalence not shown in graph).

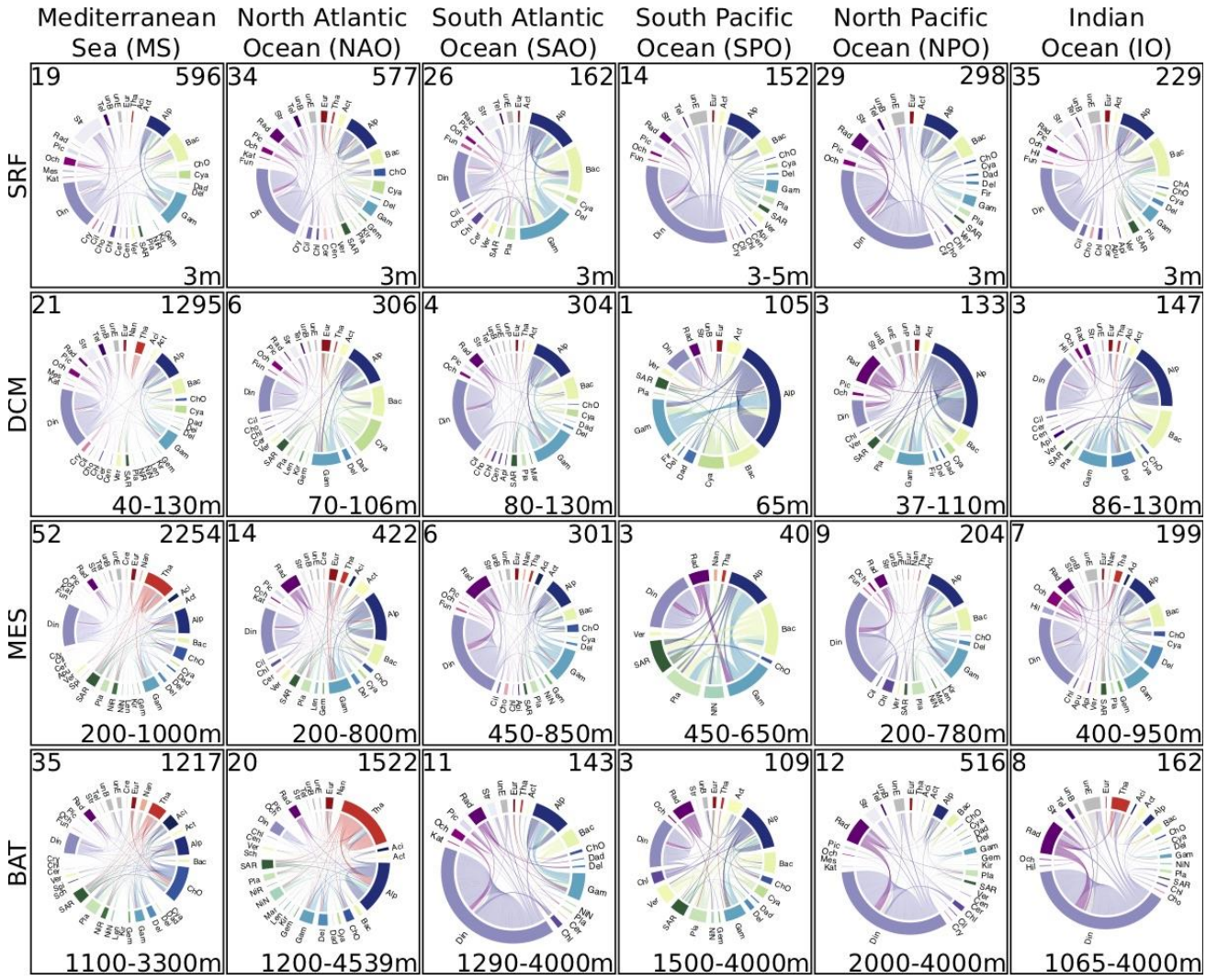


Figure 22**: Regional associations occurring in each region and depth layer**. Within a particular depth layer, if an association appears in at least one subnetwork in one region (present) and in no subnetwork in other regions (absent), it is classified as regional. The four ocean layers (rows) are surface (SRF), DCM, mesopelagic (MES), and bathypelagic (BAT). The number of samples appears in the upper left corner, the number of edges in the upper right corner, and the depth range in the lower right corner (in m below surface).

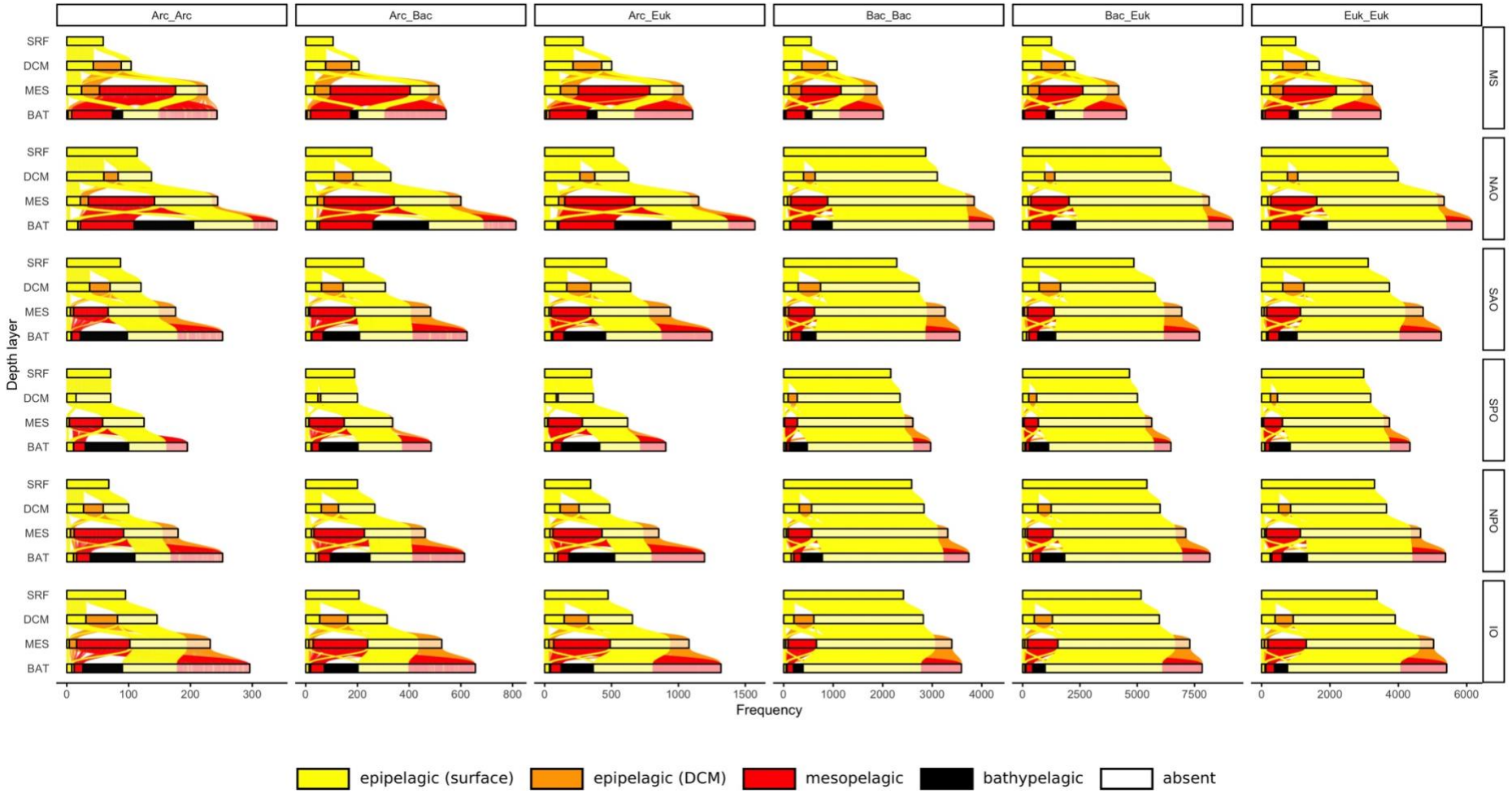


Figure 23**: Microbial associations across depth layers**. For each region and taxonomic domain, we color associations based on when they first appeared: surface (S, yellow), DCM (D, orange), mesopelagic (M, red), and bathypelagic (B, black). Absent ASVs are grouped in the white (transparent) box. Columns show associations between archaea (Arc), bacteria (Bac), and eukaryotes (Euk)

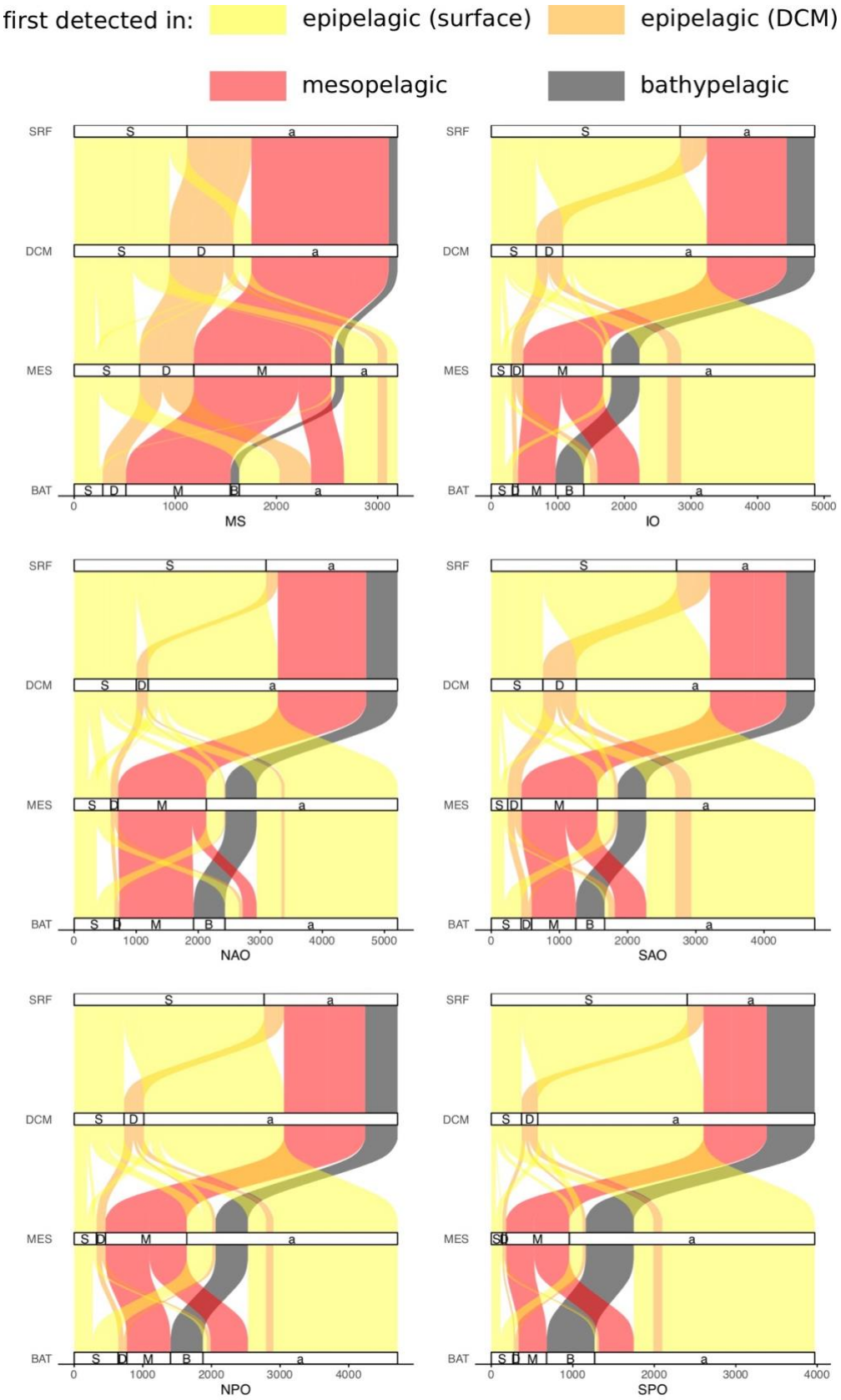


Figure 24**: ASVs across depth layers**. For each region, we color ASVs based on the layer they first appeared: surface (S, yellow), DCM (D, orange), mesopelagic (M, red), and bathypelagic (B, black). Absent ASVs are grouped in box "a". An ASV only appearing in the bathypelagic, is assigned to box "a" in above layers. That is, an ASV detected in the surface and present in the DCM but absent in lower layers, appears in the box (S) in the surface and DCM layer, but in box "a" in the meso- and bathypelagic layer. An ASV cannot be assigned to two layers. Note that most ASVs in the bathypelagic zone have been already detected in upper layers because most ASVs are assigned to the boxes "S", "D", and "M" instead of "B".

Table 17**: Fraction of microbial associations across depth layers.** For each region and layer (rows), we determined the constitution of associations (in percentage %) classifying them based on their first appearance (columns): surface, DCM, mesopelagic, and bathypelagic. We indicated the fractions above 40% in grey.

Region	Layer	Surface	DCM	Mesopelagic	Bathypelagic
MS	SRF	100.00			
	DCM	45.14	54.86		
	Mesopelagic	10.35	18.42	71.24	
	Bathypelagic	2.73	5.12	69.71	22.44
NAO	SRF	100.00			
	DCM	68.30	31.70		
	Mesopelagic	11.64	6.59	81.77	
	Bathypelagic	11.62	1.35	43.49	43.54
SAO	SRF	100.00			
	DCM	45.08	54.92		
	Mesopelagic	6.15	8.50	85.35	
	Bathypelagic	12.22	6.30	26.97	54.61
SPO	SRF	100.00			
	DCM	50.07	49.93		
	Mesopelagic	6.44	2.66	90.90	
	Bathypelagic	9.81	3.32	14.15	72.71
NPO	SRF	100.00			
	DCM	54.23	45.77		
	Mesopelagic	8.33	6.06	85.61	
	Bathypelagic	17.46	5.34	19.92	57.28
IO	SRF	100.00			
	DCM	39.23	60.77		
	Mesopelagic	5.92	7.87	86.21	
	Bathypelagic	11.00	3.84	29.61	55.56

Comparing subnetworks

Vertical and horizontal spatial variability is expected to affect network topology via biotic and abiotic variables as well as through dispersal processes (e.g., dispersal limitation). Yet, we have a limited understanding on how much marine microbial networks change due to these processes, thus analyzing the topology of subnetworks from specific ocean regions and depths is a first step to address this question. We compared the subnetworks of the six regions and depth layers using eight global network metrics (see Methods). We found that global network metrics change along the water column (Figure 25). As a general trend, subnetworks from deeper zones were more clustered (transitivity) with higher average path length, stronger associations (average positive association scores) and lower assortativity (based on degree) compared to those in surface waters. Most DCM and bathypelagic subnetworks had highest connectivity (edge density) (Figure 25).

To avoid predefined grouping into regions and depth layers, we grouped similar subnetworks via a local network metric (see Methods) and identified 36 clusters of 5 to 28 subnetworks (Table 18). We found 13 (36.1%) clusters that were dominated by surface subnetworks: six clusters (100% surface subnetworks) from three to five oceans but not MS and seven clusters with 55-86% surface networks from two to five of the six ocean regions. In turn, 11 clusters were dominated by a deeper layer: two DCM (64-90%), five mesopelagic (62-83%) and four bathypelagic dominated clusters (60-69%). Nine of these 11 clusters combined different regions except for one mesopelagic and one bathypelagic dominated cluster representing exclusively the MS

(Table 18). Furthermore, we found 11 clusters containing exclusively or mainly MS subnetworks in contrast to only one cluster dominated by an ocean basin (NAO).

Next, we built a more comprehensive representation of network similarities between subnetworks via a minimal spanning tree (MST, see Methods) to underline the pervasive connectivity of associations across depth and environmental gradients. The depth layers, ocean regions, location of clusters, and environmental factors were projected onto the MST (Figure 26). Most surface subnetworks were centrally located, while subnetworks from other depths appeared in different MST areas. Most MS subnetworks were located in a specific branch of the MST, while the five oceans were mixed, indicating homogeneity within oceans but network-based differences between the oceans and the MS. However, subnetworks in the MST tended to connect to subnetworks from the same depth layer, cluster or similar environmental conditions. All in all, the above results suggest a strong influence of environmental gradients in shaping network topology and plankton associations, as previously observed in epipelagic communities at global scale (Chaffron *et al.*, 2020).

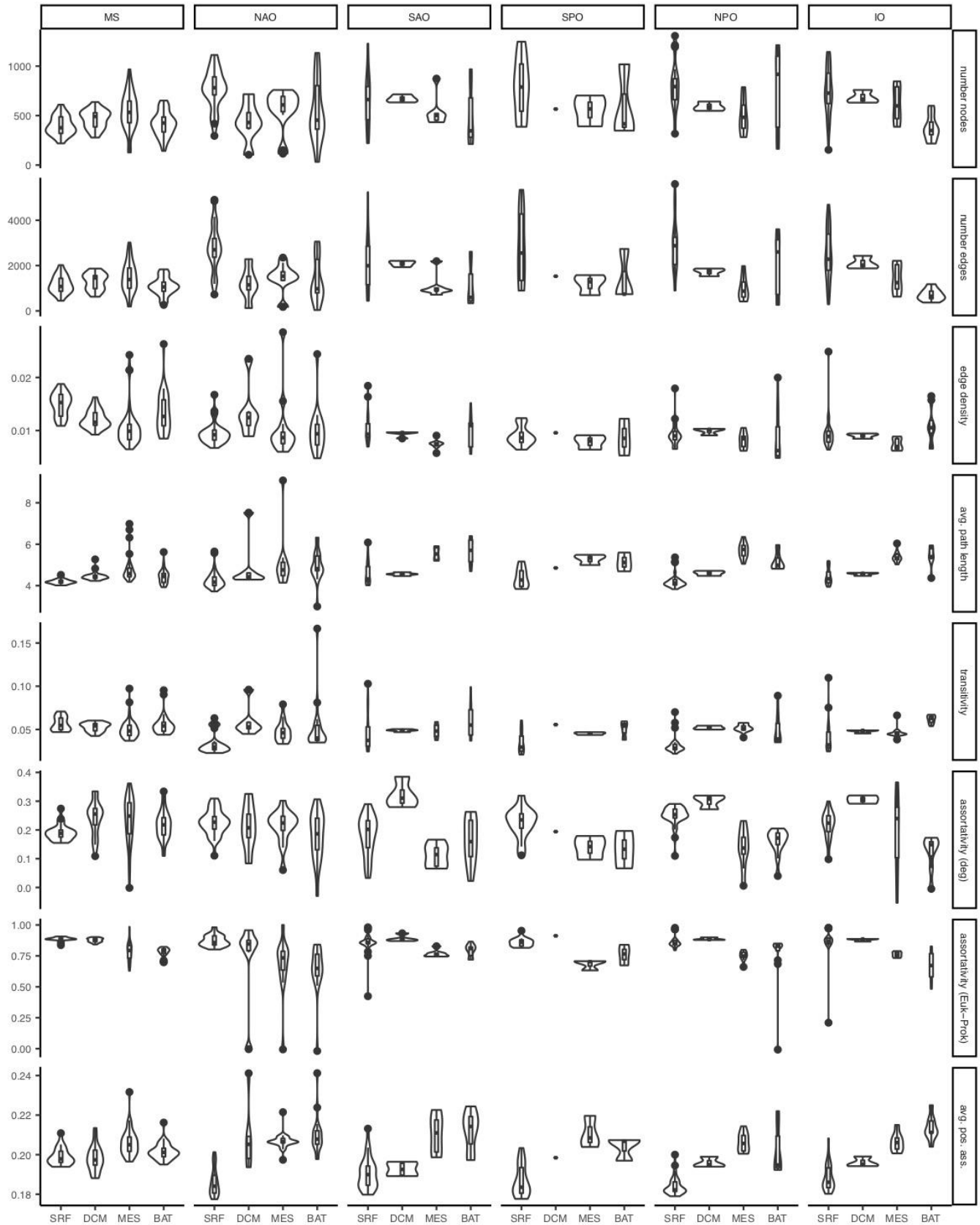


Figure 25**: Global network metrics.** The considered global network metrics are (from top to bottom): number of nodes and edges, edge density, average path length, transitivity, assortativity (degree), assortativity (eukaryote – prokaryote), and average positive association score. We grouped the metrics by region and depth layer. For better visualization, we removed one edge-density outlier: 0.07 for a bathypelagic subnetwork in the NAO (MalaVP_DNA_D2795_4000m).

Table 18: **Subnetwork cluster.** We highlighted the clusters that were dominated, i.e., over 50%, by one layer or one region in gray. The last row shows unassigned subnetworks.

Cluster ID	Dominated by	Size	Fraction of depth layers					Number of regions (if no number if indicated, it is 1x)				
			Epipelagic			Meso-pelagic	Bathy-pelagic	Epipelagic			Meso-MES	Bathy-BAT
			SRF	EPI	DCM			pelagic	EPI	DCM		
1	MS	5	20.00	20.00	20.00	20.00	20.00	SAO	MS	NAO	MS	MS
2	MS	10	10.00	-	20.00	20.00	50.00	MS	-	2xMS	2xMS	5xMS
3	MS	8	12.50	-	-	25.00	62.50	SRF	-	-	2xMS	5xMS
4	MS, MES	8	-	12.50	-	75.00	12	-	MS	-	6xMS	MS
5	MS, MES	12	16.67	-	-	66.67	16.67	IO, NAO	-	-	7xMS, NAO	2xNAO
6		8	12.50	25.00	12.50	25.00	25.00	IO	MS, NAO	NPO	MS, NAO	2xMS
7	BAT	15	13.33	-	-	26.67	60.00	IO, SPO	-	-	IO, MS, SAO, SPO	IO, MS, NAO, 2xNPO, 2xSAO, 2xSPO
8	DCM	10	10.00	-	90.00	-	-	NPO	-	5xMS, NPO, 3xSAO	-	-
9	DCM	11	36.36	-	63.64	-	-	2xNAO, NPO, SAO	-	3xIO, 2xMS, NPO, SAO	-	-
10		12	-	-	8.33	50.00	41.67	-	-	NAO	IO, MS, NAO, 2xNPO, SAO	IO, 2xNAO, NPO, SAO
11	MES	6	-	-	-	83.33	16.67	-	-	-	IO, MS, NPO, 2xSAO	IO
12	NAO, MES	6	16.67	-	-	83.33	-	NAO	-	-	2xMS, 3xNAO	-
13	SRF	11	54.55	9.09	-	27.27	9.09	IO, MS, NPO, 3xSAO	MS	-	2xMS, NAO	MS
14	BAT	16	12.50	6.25	6.25	6.25	68.75	MS, NAO	MS	MS	MS	5xNAO, 3xNPO, 2xSAO, SPO
15	SRF	8	100.00	-	-	-	-	3xIO, 4xNAO, NPO	-	-	-	-
16	MS, SRF	7	71.43	14.29	-	14.29	-	4xMS, NPO	MS	-	MS	-
17	MS	9	-	11.11	33.33	22.22	33.33	-	MS	MS, NAO, SPO	2xMS	3xMS
18	MS, BAT	8	12.50	25.00	-	-	62.50	IO	2xMS	-	-	3xMS, 2xNAO
19	SRF	7	85.72	14.29	-	-	-	2xIO, NAO, NPO, 2xSAO	MS	-	-	-
20	SRF	15	73.33	-	6.67	6.67	13.33	2xIO, 2xNAO, NPO, 5xSAO, SPO	-	MS	IO	IO, NPO
21		8	25.00	-	12.50	25.00	37.50	IO, SPO	-	MS	MS, SAO	IO, 2xNAO
22		17	23.53	-	5.88	35.29	35.29	3xSAO, SPO	-	MS	NAO, 2xNPO, SAO, 2xSPO	IO, MS, NAO, 3xSAO
23	SRF	8	75.00	12.50	-	12.50	-	IO, 2xMS, NAO, NPO, SPO	MS	-	MS	-
24	MS, MES	13	15.38	7.69	-	61.54	15.38	2xMS	MS	-	IO, 4xMS, 3xNAO	NAO, NPO
25		14	28.57	7.14	14.29	7.14	42.86	2xMS, 2xNAO	MS	2xMS	NAO	MS, 3xNPO, 2xSAO
26	SRF	7	85.72	14.29	-	-	-	2xIO-SRF, MS-EPI, 2xNAO-SRF, 2xNPO-SRF	2xIO-SRF, MS-EPI, 2xNAO-SRF, 2xNPO-SRF	-	-	-
27	SRF	11	100.00	-	-	-	-	2xIO, NAO, 4xNP, 4xSPO	-	-	-	-
28	MS	11	9.09	27.27	-	36.36	27.27	MS	3xNAO	-	4xMS	3xMS
29		12	50.00	-	16.67	16.67	16.67	IO, MS, 3xNAO, SAO	-	MS, NAO	2xMS	2xMS
30		6	50.00	-	16.67	16.67	16.67	IO, NAO, SPO	-	MS	NPO	IO-BAT
31	MS	28	25.00	10.71	7.14	35.71	21.43	4xIO, 2xMS, SAO	3xMS	2xMS	6xMS, 2xNAO, 2xNPO	IO, 2xMS, 3xNAO
32	SRF	6	100.00	-	-	-	-	IO, 2xNA, NPO, 2xSAO	-	-	-	-
33	SRF	6	100.00	-	-	-	-	NAO, 3xNPO, SAO, SPO	-	-	-	-
34	SRF	14	100.00	-	-	-	-	IO, 4xNAO, 5xNPO, 2xSAO, 2xSPO	-	-	-	-
35	SRF	13	69.23	7.69	-	-	23.08	4xIO, 3xNAO, SAO, SPO	MS	-	-	3xMS
36	SRF	7	100.00	-	-	-	-	3xIO, 3xNPO, SAO	-	-	-	-
-		24	41.67	-	12.50	29.17	16.67	2xIO, MS, 2xNAO, 3xNPO, 2xSAO	-	MS, 2xNAO	2xIO, 4xMS, NPO	MS, NAO, NPO, SAO

MS – Mediterranean Sea, NAO – North Atlantic Ocean, SAO – South Atlantic Ocean, SPO – South Pacific Ocean, NPO – North Pacific Ocean, IO – Indian Ocean, EPI – epipelagic layer, SRF – surface, DCM – Deep Chlorophyll Maximum, MES – mesopelagic layer, BAT – bathypelagic layer

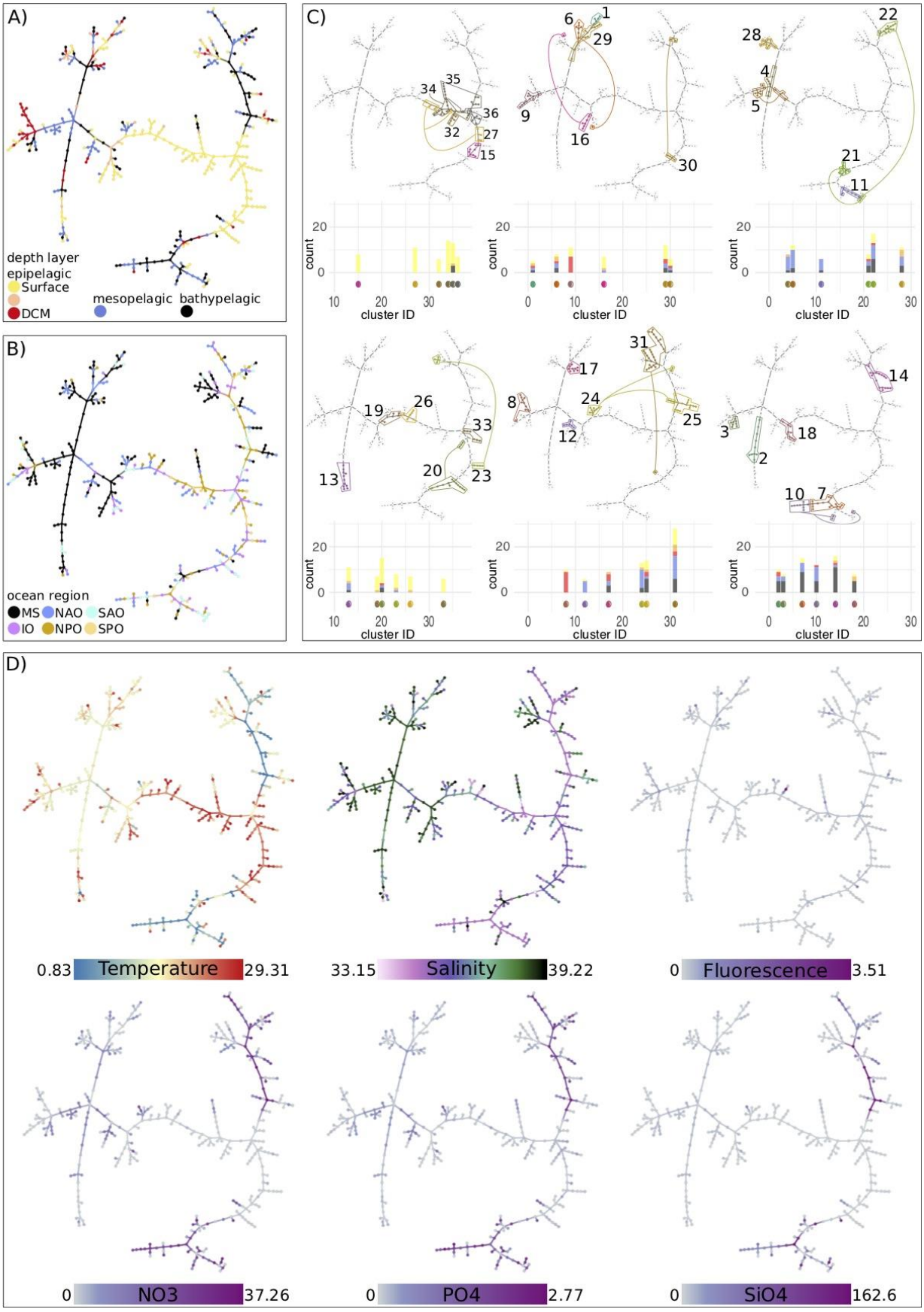


Figure 26**: Minimal Spanning Tree.** Each subnetwork is a node in the MST and represents a sample. Nodes are colored according to A) the sample's depth layer, B) the samples ocean region, C) the subnetworks cluster, and D) selected samples' environmental factors. In C), the barplots indicate the different layers within each cluster colored as in A).

Discussion

In this work, we disentangled and analyzed global and regional microbial associations across the oceans' vertical and horizontal dimensions. We found a low number of global associations indicating a potentially small global core interactome within each depth layer across six oceanic regions. Core microorganisms are often defined as those appearing in most or all samples from similar habitats (Shade & Handelsman, 2012). We previously identified a core microbiota in a coastal MS observatory based on both association patterns (Krabberød *et al.*, 2021) and temporal recurrence of associations (Deutschmann *et al.*, 2021). Both studies indicate more robust microbial connectivity, suggesting a broader core, in colder than in warmer seasons. In contrast, within each region, we found less highly prevalent associations in the bathypelagic zone of the global ocean (pointing to a smaller regional core) than in upper layers, except from the NPO, having less highly prevalent associations in the meso- than in the bathypelagic. In agreement, we found more regional bathypelagic associations than in upper layers. Thus, associations may reflect the heterogeneity and isolation of the deep ocean regions due to deep currents, water masses, or the topography of the seafloor that may prevent microbial dispersal. Moreover, the higher complexity of the deep ocean ecosystem may provide a higher number of ecological niches potentially resulting in more regional associations and agreeing with our observations. A high diversification of niches may be associated to different quality and types (labile, recalcitrant, etc.) of organic matter reaching the deep ocean from the epipelagic zone (Arístegui *et al*., 2009), which are significantly different across oceanic regions (Hansell & Carlson, 1998). In an exploration of generalists versus specialist prokaryotic metagenome-assembled genomes (MAGs) in the arctic Ocean, most of the specialists were linked to mesopelagic samples indicating that their distribution was uneven across depth layers (Royo-Llonch *et al*., 2020). This is in agreement with putatively more niches in the deep ocean than in upper ocean layers leading to more specialist taxa and subsequently more regional associations.

Vertical connectivity in the ocean microbiome is partially modulated by surface productivity through sinking particles (Mestre *et al.*, 2018; Boeuf *et al.*, 2019; Ruiz-González *et al.*, 2020). An analysis of eight stations, distributed across the Atlantic, Pacific and Indian oceans (including 4 depths: Surface, DCM, meso- and bathypelagic), indicated that bathypelagic communities comprise both endemic taxa as well as surface-related taxa arriving via sinking particles (Mestre *et al.*, 2018). Ruiz-González *et al.* (Ruiz-González *et al.*, 2020) identified for both components (i.e., surface-related and deep-endemic) the dominating phylogenetic groups: while *Thaumarchaeota*, *Deltaproteobacteria*, *OM190* (*Planctomycetes*) and *Planctomycetacia* (*Planctomycetes*) dominated the endemic bathypelagic communities, *Actinobacteria*, *Alphaproteobacteria*, *Gammaproteobacteria* and *Flavobacteriia* (*Bacteroidetes*) dominated the surface-related taxa in the bathypelagic zone. We found association partners for each dominating phylogenetic group within each investigated type of association, i.e., highly prevalent, regional,

global, prevalent, and low-frequency associations. While ASVs belonging to these taxonomic groups were present throughout the water column, specific associations were observed especially in the mesopelagic and the bathypelagic zones, which suggests specific associations between deep-sea endemic taxa. This is in agreement with a recent study that found a remarkable taxonomic novelty in the deep ocean by analyzing 58 microbial metagenomes from global samples, unveiling ~68% archaea and ~58% bacterial novel species (Acinas *et al.*, 2021).

Less is known about associations found along the entire or a substantial fraction of the water column, suggesting consortia of associated microorganisms that sink together or that populate large vertical ranges of the water column. Associations present across all layers were few but may represent interacting taxa that populate the entire water column or that sink together. However, given that we targeted mainly picoplankton, we would not expect a considerable influence of sinking particles in the vertical distribution of associations in this study. Some associations observed in the deep ocean may correspond to consortia of taxa degrading sinking particles, or taxa that might have detached from sinking particles, i.e., dual life-style taxa as observed in (Sebastián, Sánchez, *et al.*, 2021). Alternatively, microorganisms may have reached bathypelagic waters via fast-sinking processes, embedded in (larger) particles (Agusti *et al.*, 2015). By following this observation, a previous study found that the abundances of microorganisms in deeper layers mirrored the changes in abundance of microorganisms in shallower layers, at a single sampling station, indicating that communities populating different ocean depths are not isolated from each other but linked, possibly through sinking particles or migrating organisms transporting nutrients through the water column (Cram, Xia, *et al.*, 2015). However, microbial co-occurrence alone does not suffice to infer microbial interactions, because different mechanisms, such as selection or dispersal, influence species as well as their interactions (Poisot *et al.*, 2012). Our results suggest that microorganisms can potentially change their interaction partners along with vertical (and horizontal) scales and, to a lesser extent, maintain interactions along the water column.

A study of global-ocean picoplanktonic eukaryotes through the water column (from the epi- to the bathypelagic zone) found the highest and lowest relative metabolic activity for most eukaryotes in the meso- and bathypelagic zones, respectively (Giner *et al.*, 2020). Thus, we could hypothesize more competition in the mesopelagic zone and more beneficial interactions in the bathypelagic zone. In our study, mesopelagic subnetworks displayed the lowest connectivity in most regions on average, and we found the strongest associations among both meso- and bathypelagic subnetworks. Moreover, we found the highest clustering (transitivity) in the meso- and bathypelagic zones (relatively colder waters) compared to the epipelagic zone (warmer waters). Similarly, a previous global-scale study (Chaffron *et al.*, 2020) concentrating on the epipelagic zone and including polar waters, found higher edge density, association strength and clustering in polar (colder waters) compared to warmer waters. These results suggest that either

microorganisms interact more in colder and darker environments or that their recurrence is higher due to a higher environmental selection exerted by low temperatures and no light. Alternatively, limited resources (primarily nutrients) in the surface versus deep ocean may prevent the establishment of specific microbial interactions. Furthermore, another explanation could be the higher diversity of ecological niches and, thus, a higher diversity of associations in the meso- and bathypelagic.

Through quantifying regional associations, our results indicated distinct associations in the MS, where most regional associations were observed compared to the global ocean, as previously shown in an epipelagic network (Lima-Mendez *et al.*, 2015). Furthermore, we found a substantial number of regional associations in the NAO compared to other ocean basins, contrasting with the NAO having the lowest number of regional associations in a previous epipelagic network (Lima-Mendez *et al.*, 2015).

Conclusion

Our network-based exploration disentangles the spatial distribution of associations of the global ocean microbiome, from top to bottom layers, suggesting both global and regional interactions. Our analysis demonstrated the change of network topology across vertical (water column) and horizontal (different regions) dimensions of the ocean. Furthermore, our results indicate that associations have specific spatial distributions that are not just mirroring ASV distributions.

Methods

Dataset

Samples originated from two expeditions, Malaspina-2010 (Duarte, 2015) and Hotmix (Martínez-Pérez *et al.*, 2017). The former was onboard the R/V Hespérides and most ocean basins were sampled between December 2010 and July 2011. Malaspina samples included i) *MalaSurf*, surface samples (Ruiz-González *et al.*, 2019; Logares *et al.*, 2020), ii) *MalaVP*, vertical profiles (Giner *et al.*, 2020), and iii) *MalaDeep*, deep-sea samples as in (Pernice *et al.*, 2016; Salazar *et al.*, 2016; Sanz-Sáez, 2021). For the Hotmix expedition, sampling took place onboard the R/V Sarmiento de Gamboa between 27th April and 29th May 2014 and represented a quasi-synoptic transect across the MS and the adjacent North-East of the NAO. See details in Table 19.

DNA extractions are indicated in the papers associated with each dataset (Table 19). From the DNA extractions, the 16S and 18S rRNA genes were amplified and sequenced. PCR amplification and sequencing of *MalaSurf*, *MalaVP* (18S), and *Hotmix* (16S) are indicated in the papers associated with each dataset in Table 19. *MalaVP* (16S) and *Hotmix* (18S) were PCR-amplified and sequenced following the same approach as in (Logares *et al.*, 2020). *MalaDeep* samples were obtained from (Pernice *et al.*, 2016; Salazar *et al.*, 2016) but re-sequenced in

Genoscope (France) with different primers, as described below. *MalaSurf*, *MalaVP* and *Hotmix* datasets were sequenced at RTL Genomics (Texas, USA).

Table 19**: Dataset compilation**. Our data was a compilation of four different datasets. We required that each location had to provide data for both eukaryotes and prokaryotes, which resulted in 397 samples. This condition allowed only 13 MalaDeep samples.

Dataset	Samples used for analysis	Stations	Depth range (m)	Water samples	Size Fraction (µm)	16S	18S	Reference	ENA accession number
Malaspina									
MalaSurf	122	120	3	122	0.2-3	122	124	(Ruiz-González et al., 2019; Logares et al., 2020)	PRJEB23913 [18S rRNA genes], PRJEB25224 [16S rRNA genes]
MalaVP	83	13	3-4000	91	0.2-3	91	83	(Giner et al., 2020) & This study	PRJEB23771 [18S rRNA genes], PRJEB45015 [16S rRNA genes]
MalaDeep (Prok)	13	30	~4000	60	0.2-0.8	41	-	(Sanz-Sáez, 2021)	PRJEB45011
MalaDeep (Euk)	13	27	2400-4000	27	0.8-20	-	82	This study	PRJEB45014
Hotmix	179	29	3-4539	188	0.2-3	188	179	(Sebastián, Ortega-Retuerta, *et al.*, 2021)	PRJEB44683 [18S rRNA genes], PRJEB44474 [16S rRNA genes]

16S and 18S refer to sequenced samples; Prok - prokaryotes; Euk – eukaryotes

We used the same amplification primers for all samples. For the 16S, we amplified the V4-V5 hypervariable region using the primers 515F-Y and 926R (Parada *et al.*, 2016). For the 18S, we amplified the V4 hypervariable region with the primers TAReukFWD1 and TAReukREV3 (Stoeck *et al.*, 2010). See more details in (Logares *et al.*, 2020). Amplicons were sequenced in *Illumina* MiSeq or HiSeq2500 platforms (2x250 or 2x300 bp reads). Inference of Amplicon Sequence Variants (ASVs) was made using DADA2 (Callahan *et al.*, 2016), v.1.4.0, running each dataset separately before merging the results. ASVs were assigned taxonomy using SILVA (Quast *et al.*, 2012), v132, for prokaryotes, and PR2 (Guillou *et al.*, 2012), v4.11.1, for eukaryotes. ASVs corresponding to Plastids, Mitochondria, Metazoa, and Plantae, were removed. Only samples with at least 2000 reads were kept. The dataset contained several *MalaDeep* replicates, which we merged, and two filter sizes: given the cell sizes of prokaryotes versus microeukaryotes, we selected the smallest available filter size (0.2-0.8 µm) for prokaryotes and the larger one (0.8-20 µm) for microeukaryotes. The other three datasets used the filter sizes of 0.2-3 µm. Additionally, we required that samples had eukaryotic and prokaryotic data, resulting

in 397 samples for downstream analysis: 122 *MalaSurf*, 83 *MalaVP*, 13 *MalaDeep*, and 179 *Hotmix*. We separated the samples into epipelagic, mesopelagic and bathypelagic zone (Figure 1). Furthermore, we separated most epipelagic samples into surface and deep-chlorophyll maximum (DCM) samples, but 18 MS and 4 NAO samples belonged to neither. We also considered environmental variables: Temperature (2 missing values = mv), salinity (2 mv), fluorescence (3 mv), and inorganic nutrients NO_3^- (36 mv), PO_4^{3-} (38 mv), and SiO_2 (37 mv), which were measured as indicated elsewhere (Giner *et al.*, 2020; Logares *et al.*, 2020; Sebastián, Ortega-Retuerta, *et al.*, 2021). In specific samples, missing data on nutrient concentrations were estimated from the World Ocean Database (Boyer *et al.*, 2013).

Single static network

We constructed the single static network in four steps. First, we prepared the data for network construction. We excluded rare microorganisms by keeping ASVs with a sequence abundance sum above 100 reads and appearing in at least 20 samples (>5%). The latter condition removes bigger eukaryotes only appearing in the 13 *MalaDeep* eukaryotic samples of a bigger size fraction. To control for data compositionality (Gloor *et al.*, 2017), we applied a centered-log-ratio transformation separately to the prokaryotic and eukaryotic tables before merging them.

Second, we inferred a (preliminary) network using FlashWeave (Tackmann *et al.*, 2019), selecting the options “heterogeneous” and “sensitive”. FlashWeave was chosen as it can handle sparse datasets like ours, taking zeros into account and avoiding spurious correlations between ASVs that share many zeros.

Third, we aimed to remove environmentally-driven edges. FlashWeave (Tackmann *et al.*, 2019) could detect indirect edges and allows to supply additional metadata such as environmental variables, but currently does not support missing data. Thus, we applied EnDED (Deutschmann et al. 2020), combining the methods Interaction Information (with 0.05 significance threshold and 10000 iterations) and Data Processing Inequality as done previously via artificially-inserted edges to connect all microbial nodes to the six environmental parameters (Deutschmann *et al.*, 2021). Although EnDED can handle missing environmental data when calculating intermediate values relating ASV and environmental factors, it would compute intermediate values for microbial edges using all samples. Thus, to avoid a possible bias and speed up the calculation process, we applied EnDED individually for each environmental factor, using only the samples containing values for the specific environmental factor.

Fourth, we removed isolated nodes, i.e., nodes without any edge. The resulting network represented the single static network in our study.

Sample-specific subnetwork

We constructed 397 sample-specific subnetworks. Each subnetwork represented one sample and was derived from the single static network, i.e., a subnetwork contained nodes and edges present in the single static network but not vice versa. Consider sample s_{RL} with R being the marine region, and L is the sample's depth layer. Let e be an association between microorganisms A and B. Then, association e is present in the sample-specific subnetwork N_s, if

i. e is an association in the single static network,
ii. the microorganisms A and B are present within sample s, i.e., the abundances are above zero within that particular sample, and
iii. the association has a region and depth specific Jaccard index, J_{RL}, above 20% (see below).

In addition to these three conditions, a node is present in a sample-specific subnetwork when connected to at least one edge, i.e., we removed isolated nodes.

Regarding the third condition, we determined J_{RL} for each association pair by computing within each region and depth layer, the fraction of samples two microorganisms appeared together (intersection) from the total samples at least one microorganism appears (union). Table 20 shows the number of edges using different thresholds. Given the heterogeneity of the dataset within regions and depth layers, we decided on a low threshold, keeping edges with a Jaccard index above 20% and removed edges below or equal to 20%. We tested the robustness by randomly drawing a subset of samples of each region and depth combination. The subset contained between 10% and 90% of the original samples. We rounded up decimal numbers to avoid zero samples, e.g., 10% of 4 or 7 samples both resulted in a subset of 1 sample. We excluded DCM of the SPO because it contained only one sample. Next, we recomputed the Jaccard index for the random subset. Lastly, requiring J>20%, we evaluated the robustness determining i) how many edges were kept in the random subsamples compared to all samples, and ii) how many edges were kept in the random subset that were also kept when all samples were used. We repeated the procedure for each region-depth combination 1000 times.

Table 20**: Different thresholds on the Jaccard index.** Number of edges within each region and depth layer before (J>0%) and after filtering edges with low Jaccard index measuring how often the association partners appeared together in the region and depth layer. The DCM layer in the South Pacific Ocean (SPO) contained only one subnetwork, which resulted in the edge prevalence being 100% for all edges.

Region	Layer	Samples	Depth (m)	J>0%	J>10%	J>20%	J>30%	J>40%	J>50%
MS	EPI - SRF	19	3	3710	3631	3263	2881	2375	1797
	EPI	18	12-50	4763	4682	4196	3731	3064	2189
	EPI - DCM	21	40-130	5545	5417	4736	4030	3062	2027
	MES	52	200-1000	8756	8403	7336	6179	4629	3088
	BAT	35	1100-3300	4497	4263	3694	3171	2506	1830
NAO	EPI - SRF	34	3	15862	15255	13478	11449	8487	5331
	EPI	4	50	3027	3027	3027	2778	2529	2091
	EPI - DCM	6	70-106	3865	3865	3738	3480	2973	2212
	MES	14	200-800	6325	6289	5689	5109	4169	2978
	BAT	20	1200-4539	7490	7419	6831	6206	5211	3857
SAO	EPI - SRF	26	3	13118	12768	11026	9269	6842	4353
	EPI - DCM	4	80-130	4199	4199	4199	3941	3443	2468
	MES	6	450-850	3937	3937	3740	3440	2687	1614
	BAT	11	1290-4000	4143	4130	3886	3605	3049	2254
NPO	EPI - SRF	29	3	14376	13778	11919	9907	7323	4736
	EPI - DCM	3	37-110	3100	3100	3100	3100	2568	1968
	MES	9	200-780	4197	4197	3781	3343	2583	1625
	BAT	12	2000-4000	5198	5185	4834	4510	4009	3372
SPO	EPI - SRF	14	3-5	12007	11927	10420	8990	6728	4480
	EPI - DCM	1	65	1530	1530	1530	1530	1530	1530
	MES	3	450-650	2066	2066	2066	2066	1756	1318
	BAT	3	1500-4000	3159	3159	3159	3159	2906	2128
IO	EPI - SRF	35	3	14307	13646	11736	9602	6912	4396
	EPI - DCM	3	86-130	3411	3411	3411	3411	2855	2310
	MES	7	400-950	4654	4654	4344	3961	3083	2082
	BAT	8	1065-4000	2928	2928	2790	2563	2101	1290

MS – Mediterranean Sea, NAO – North Atlantic Ocean, SAO – South Atlantic Ocean, SPO – South Pacific Ocean, NPO – North Pacific Ocean, IO – Indian Ocean, EPI – epipelagic layer, SRF – surface, DCM – Deep Chlorophyll Maximum, MES – mesopelagic layer, BAT – bathypelagic layer

Spatial recurrence

To determine an association's spatial recurrence, we calculated its prevalence as the fraction of subnetworks in which the association was present. We determined association prevalence across the 397 samples and each region-layer combination. We mapped the scores onto the single static network, visualized in Gephi (Bastian *et al.*, 2009), v.0.9.2, using the Fruchterman Reingold Layout (Fruchterman & Reingold, 1991) with a low gravity score of 0.5. We used the region-layer prevalence to determine global and regional associations. We considered an association to be global within a specific depth layer if its prevalence was above 70% in all regions. In turn, a regional association had an association prevalence above 0% within a particular region-layer (present, appearing in at least one subnetwork) and 0% within other regions of the same layer (absent, appearing in no subnetwork). In addition, associations that are not global but appear in all regions over 50% are considered prevalent. Similarly, associations that are not global nor prevalent but appear in all regions over 20% are considered low-frequency. Thus, an association can be classified as i) global, ii) regional, iii) prevalent, iv) low-frequency, and v) "other", i.e., associations that have not been classified into the previous categories.

Global network metrics

We considered the *number of nodes* and *edges* and six other global network metrics of which most were computed with functions of the igraph R-package (Csardi & Nepusz, 2006). *Edge density* indicating connectivity is computed through the number of actual edges divided by the number of possible edges. The *average path length* is the average length of all shortest paths between nodes in a network. *Transitivity* indicating how well a network is clustered is the probability that the nodes' neighbors are connected. *Assortativity* measures if similar nodes tend to be connected, i.e., *assortativity (degree)* is positive if high degree nodes tend to connect to other high degree nodes and negative otherwise. Similarly, *assortativity (Euk-Prok)* is positive if eukaryotes tend to connect to other eukaryotes and prokaryotes tend to connect to other prokaryotes. Lastly, we computed the *average positive association strength* as the mean of all positive association scores provided by FlashWeave (Tackmann *et al.*, 2019).

Local network metric

The previous global metrics disregard local structures' complexity, and topological analyses should include local metrics (Espejo *et al.*, 2020), e.g., graphlets (Pržulj *et al.*, 2004). Here, we determined network-dissimilarity between each pair of sample-specific subnetworks proposed in (Yaveroğlu *et al.*, 2014), comparing network topology without considering specific ASVs. The network-dissimilarity is a distance measurement that is always positive: 0 if networks are identical and greater numbers indicate greater dissimilarity.

Next, we constructed a Network Similarity Network (NSN), where each node is a subnetwork and each node connects to all other nodes, i.e., the NSN was a complete graph. We assigned the network-dissimilarity score as edge weight within the NSN. To simplify the NSN while preserving its main patterns, we determined the minimal spanning tree (MST) of the NSN. The MST had 397 nodes and 396 edges. The MST is a backbone, with no circular path, in which the edges are chosen so that the edge weights sum is minimal and all nodes are connected, i.e., a path exists between any two nodes. We determined the MST using the function *mst* in the igraph package in R (Prim, 1957; Csardi & Nepusz, 2006).

Using the network-dissimilarity (distance) matrix, we determined clusters of similar subnetworks in python. First, we reduced the matrix to ten dimensions using *umap* (McInnes *et al.*, 2018) with the following parameter settings: n_neighbors=3, min_dist=0, n_components=10, random_state=123, and metric='precomputed'. Second, we clustered the subnetworks (represented via ten dimensions) with *hdbscan* (McInnes *et al.*, 2017) setting the parameters to min_samples=3 and min_clusters=5.

Reproducibility

R-Markdowns for data analysis including commands to run FlashWeave and EnDED (environmentally-driven-edge-detection and computing Jaccard index) are publicly available: https://github.com/InaMariaDeutschmann/GlobalNetworkMalaspinaHotmix. While the networks are already available, the microbial sequence abundances (ASV table), taxonomic classifications, environmental data including nutrients will be publicly available after acceptance. The data are of course available upon request to reviewers.

Supplementary Material

Due to their size, the two supplementary materials are not included in this thesis but publicly available as tab-separatec txt-files in the Github repository (specific links are provided below): https://github.com/InaMariaDeutschmann/GlobalNetworkMalaspinaHotmix.

Supplementary Material 1: Highly prevalent (>70%) regional associations. For each association between two ASVs (first and second column) we list: region (third column), depth layer (fourth column), prevalence in that region and depth layer (fifth column), type: eukaryotic (Euk_Euk), prokaryotic (Prok_Prok), and association between domains (Euk_Prok) (sixth column), and the phyla (seventh and eight column).
Available here:
https://github.com/InaMariaDeutschmann/GlobalNetworkMalaspinaHotmix/blob/main/05_ClassifyingAssociations/HighlyPrevalentRegionalAssociations.txt

Supplementary Material 2: Associations appearing in all layers in at least one region. For each association between two ASVs (first and second column) we list: the classification in each layer (3-6 column), overall prevalence (8. column), prevalence in each region and depth layer (9- 34. column), the number of regions in which the association appeared in all layers (AllLayers, 35. column), the number of layers an association appears in a region (36-41. column), type: eukaryotic (Euk_Euk), prokaryotic (Prok_Prok), and association between domains (Euk_Prok) (42. column), and the phyla (43-44. column).
Available here:
https://github.com/InaMariaDeutschmann/GlobalNetworkMalaspinaHotmix/blob/main/06_VerticalConnectivity/AllLayersAssociations.txt

Acknowledgements

We thank all members of the Malaspina and Hotmix expeditions with the multiple projects funding these collaborative efforts. Sampling was carried out thanks to the Consolider-Ingenio programme (project Malaspina 2010 Expedition, ref. CSD2008–00077) and HOTMIX project (CTM2011-30010/MAR), funded by the Spanish Ministry of Economy and Competitiveness Science and Innovation. Part of the analyses have been performed at the Marbits bioinformatics core at ICM-CSIC (https://marbits.icm.csic.es).

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 (ESR2, http://www.singek.eu) to RL. 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), MicroEcoSystems (240904, RCN, Norway) and MINIME (PID2019-105775RB-I00, AEI, Spain) to RL. SC was supported by the CNRS MITI through the interdisciplinary program Modélisation du Vivant (GOBITMAP grant). SC, DE and SGA were funded by the H2020 project AtlantECO (award number 862923).

Author's contributions

The overall project was conceived and designed by RL. JMG, CMD, SGA, RM, JA were responsible for the sampling and acquisition of contextual data. CRG, JP and MS processed specific samples in the laboratory. RL processed the amplicon data generating the two ASV tables. They were the starting point of the present study, which is part of the overall project. IMD developed the conceptual approach and DE, SC, and RL contributed to its finalization. IMD performed the data analysis. ED, MS, CMD, SGA, RM, JMG, DE, SC, and RL contributed with interpretation of the results. IMD wrote the original draft. All authors contributed to manuscript revisions and approved the final version of the manuscript.

Final remarks

- ⇒ Associations could be global or regional.
- ⇒ Most sampling does not allow to construct one network per location because only one sample is provided.
- ⇒ Our post-network-construction approach allows determining sample-specific subnetworks derived from an overall single static network.
- ⇒ Our approach can be used to quantify spatial recurrence for each association over the entire dataset and also for each depth or region and depth.

⇒ The fraction of global associations is highest in the DCM layer and lowest in the bathypelagic zone.

⇒ Regional associations increase with depth.

⇒ Most regional associations have a low prevalence.

⇒ High prevalent associations may infer core associations that are essential for ecosystem functioning.

⇒ Our approach allows to quantify associations based on spatial recurrence, which was suggested in Chapter 5 to may strengthen and reduce the number of potential interaction hypotheses.

⇒ Here, we accomplished the proposed idea from the previous chapter: we adjusted our approach of quantifying edge recurrence (temporal) to determine spatial recurrence.

⇒ As in the previous chapter, we inserted artificial edges (artificially-generated triplets) to apply EnDED on each edge with each possible environmental factor. The fraction of indirect edges was lower in this study considering different locations and depths, than in the previous chapter considering ten years of one coastal surface location.

⇒ As known from other works, most deep ocean ASVs already appeared in upper layers. In contrast, most associations in the meso- and bathypelagic appeared for the first time.

⇒ Mapping environmental factors onto the MST, the tendency of similar environmental variables grouping together indicates a possible connection between environment and network topology since the MST was constructed using network similarity based on solely network topology.

⇒ The greatest bias of the study is the heterogeneity of the sampling in terms of the number of samples.

Chapter 8 Further investigations including EnDED

The chapters' focus is on three aspects: i) comparing network-based environmental and artificially-generated triplets, ii) comparing the application of EnDED in networks constructed with network-construction tools used in first and co-authored manuscripts, iii) expanding the type of data used to detect factors causing indirect dependencies.

Network-based environmental versus artificially-generated triplets

Chapter 5 and Chapter 6 employ the same model marine microbial BBMO network. We used *network-based environmental triplets* in the former chapter and *artificially-generated triplets* in the latter chapter. Network-based environmental triplets are generated during network construction. However, not all methods allow including environmental factors as nodes, e.g., SparCC (Friedman & Alm, 2012). Therefore, we adjusted the method. First, we construct a network without environmental factors. Second, we insert artificial edges from each microbial node to each environmental factor, which results in artificially-generated triplets.

The network-based triplets approach requires to re-construct the network if environmental data changes, in contrast to the artificially-generated triplets approach. Both approaches can be used with the current implementation of EnDED. Although artificially-generated triplets increase computation time, the approach is highly suitable for parallelization since EnDED can be run with each environmental factor separately. Further, artificial edges do not provide information about a potential association between ASV and environmental variables, e.g., no sign of association strength and no association duration, which are used by the methods Sign Pattern and Overlap, respectively. Thus, the artificially-generated triplets approach can only use the methods Interaction Information, Data Processing Inequality, and their ensemble approach (intersection combination). However, these methods may be sufficient given that Sign Pattern and Overlap indicated all edges as environmentally-driven using network-based environmental triplets in the BBMO network (Chapter 5). Finally, the artificially-generated triplets approach applies EnDED on more edges, which may increase the chance of finding environmentally-driven edges. However, if the number of environmentally-driven edges would remain the same, it would indicate that the network-based triplets are sufficient and the computationally costlier artificially-generated triplet approach would not be needed. The results indicated that this was not the case.

The BBMO network before applying EnDED contained 29820 edges of which 2488 (8.34%) are environmentally-driven using network-based triplets, in contrast to 3315 (11.12%) when using artificially-generated triplets. Thus, adapting the strategy allowed to identify more environmentally-driven edges, especially indirect edges due to nutrients and the total chlorophyll-a concentration (Table 21). There was a small increase in identified environmentally-driven edges

due to temperature and day length (hours of light). This indicates that the chosen network construction tool may incorporate associations between microorganisms and temperature and day length well, but misses relevant associations with other factors such as nutrients. The discrepancy may be due to temperature and day-length being easy to quantify in contrast to other factors.

To conclude, EnDED allows indirect dependency detection using network-based environmental triplets (generated during network construction) and artificially-generated triplets (generated after network construction). We quantified more environmentally-driven edges using artificially-generated triplets, especially for nutrients and the total chlorophyll-a concentration but only slightly more for temperature and day-length.

Table 21: **Number and fraction of environmentally-driven edges for each available environmental factor.** We detected environmentally-driven edges with EnDED using network-based environmental triplets and artificially generated environmental triplets, i.e., we introduced artificial edges to connect each ASV with each environmental factor.

Environmental factor	Network-based triplet	Artificially generated triplet
Temperature	1831 (6.14%)	1920 (6.44%)
Total chlorophyll-a concentration	175 (0.59%)	838 (2.81%)
Day length	652 (2.19%)	730 (2.45%)
NO_2^-	0	192 (0.64%)
SiO_2	5 (0.02%)	162 (0.54%)
NO_3^-	1 (0.003%)	57 (0.19%)
Turbidity	Not in a triplet	47 (0.16%)
NH_4^+	Not in a triplet	0
PO_4^{3-}	Not in network	0
Salinity	Not in network	0

Comparing the application of EnDED on networks constructed with different tools using the Malaspina Surface data

Malaspina Surface data has been used in three projects. In the first study (Logares *et al.*, 2020), we used the method SparCC (Friedman & Alm, 2012) as implemented in FastSpar (Watts *et al.*, 2019) to construct and compare the prokaryotic and the eukaryotic network. The network construction method does not allow including environmental factors as nodes and the analysis was done before the completion of EnDED. In the second study (Latorre *et al.*, 2021), we used the network construction tool MICtools (Albanese *et al.*, 2018), which allows to include environmental data within the network construction. EnDED has been applied using the traditional network-based environmental triplet approach with the methods Interaction Information and Data Processing Inequality. This second marine microbial investigation focused on a specific microbial group. Specifically, we used an association network to infer potential interaction partners for MAST-4A/B/C/E in the sunlit global ocean. Thus, although the study aim revolved around MAST-4 taxa, a sub-aim was to provide potential interaction partners via a network-approach. To select strong interaction partner candidates several steps have been done starting with the sequence abundance table and ending with a list of interaction partners. To select

promising interaction partner candidates, the strategy comprised different filters. We observed a vast number of spurious associations in the preliminary network. Thus, we filtered associations whose association partners had many matching zeros, i.e., we removed associations with half or more zero-matching samples. Moreover, we applied a Jaccard index filter of 25%, removed environmentally-driven associations and only considered those with an association strength above 0.4. The experience from the first and second study resulted in the selection of a third network construction tool in the third study. The third project used a data compilation including Malaspina Surface data and three other datasets to cover the water column from the surface to the deep ocean (Chapter 7). The sparse sequence abundance table (many zeros) was prone to result in spurious associations. To avoid spurious associations, we used the network construction tool FlashWeave in the heterogenous mode, which ignores matching zeros when computing associations (Tackmann *et al.*, 2019).

In this chapter, we use the Malaspina Surface dataset as an example dataset to compare environmentally-driven associations in eukaryotic and prokaryotic networks constructed using the three tools employed in the three projects. Here, we use EnDED with artificially-generated triplets. Before network construction, we required that all samples contained eukaryotic and prokaryotic data resulting in 122 samples. We removed rare ASVs and kept ASVs with a sequence abundance sum above 100 reads and appearing in at least 19 samples (>15%). To control for data compositionality (Gloor *et al.*, 2017), we applied a centered-log-ratio transformation separately to the prokaryotic and eukaryotic tables before applying the tools MICtools and FlashWeave. The tool FastSpar expects raw counts.

The *SparCC network* was constructed using SparCC (Friedman & Alm, 2012) as implemented in FastSpar (Watts *et al.*, 2019), v.0.0.7. It was run with 1000 iterations including 1000 bootstraps to infer p-values resulting in the smallest being 0.001. We only considered associations with a p-value of 0.001. The *MICtools network* was constructed using Maximal Information Coefficient (MIC) analyses as implemented in MICtools (Albanese *et al.*, 2018), v.1.0.1. TICe null distributions were estimated using 200000 permutations. The significance level for the MICtools network was set to 0.001 as suggested in (Weiss *et al.*, 2016). The *FlashWeave network* was constructed using FlashWeave (Tackmann *et al.*, 2019), v.0.18.0, in Julia, v.1.5.3, selecting the heterogeneous and sensitive mode.

As expected, the MICtools networks contain a vast number of associations and FlashWeave networks considerably less associations (Table 22). The discrepancy may be due to spurious associations. Thus, we use the Jaccard index as an estimator or indicator for spurious associations. We determined the Jaccard index for each association in each of the six networks (Figure 27). The distribution of Jaccard index of same tools is similar between eukaryotic and prokaryotic networks and different between networks constructed with different tools. The distribution from SparCC networks (first row in Figure 27) is more similar to the FlashWeave

networks (third row in Figure 27); they show two peaks whereas the MICtools networks (second row in Figure 27) show one peak. Considering associations with a Jaccard index equal or below 25%, we found the highest number of spurious associations in the MICtools networks (60.4-70.8%), followed by SparCC networks (40.9-42.7%), and less spurious associations in FlashWeave networks (19.1-23.42%). For further analysis, we only keep associations with a Jaccard index above 25% and remove proportionally more associations in eukaryotic than prokaryotic networks constructed with MICtools and FlashWeave, but slightly less in SparCC networks.

Table 22: **Number and fraction of associations in the network before and after applying the Jaccard index filter and EnDED.** We used a Jaccard index threshold of 0.25, i.e., association partners have to co-occur in more than 25% of the samples in which one or both were detected. Moreover, we list the number and fraction of environmentally-driven edges for each network and available environmental factor detected with EnDED using artificially generated environmental triplets, i.e., we introduced artificial edges to connect each ASV with each environmental factor.

Domain	Eukaryotic networks			Prokaryotic networks		
and tool	SparCC	MICtools	FlashWeave	SparCC	MICtools	FlashWeave
Significant	43482	1275951	4402	19676	123599	1227
Jaccard index						
removed	17786	902745	1031	8404	74608	234
	(40.9%)	(70.8%)	(23.42%)	(42.7%)	(60.4%)	(19.1%)
kept	25696	373206	3371	11272	48991	993
	(59.1%)	(29.2%)	(76.6%)	(57.3%)	(39.6%)	(80.9%)
EnDED						
removed	1093	42464	88	517	4477	10
	(4.3%)	(11.4%)	(2.6%)	(4.6%)	(9.1%)	(1.0%)
kept	24603	330742	3283	10755	44514	983
	(95.7%)	(88.6%)	(97.4%)	(95.4%)	(90.9%)	(99.0%)
Environmental drivers						
Temperature	137 (0.5%)	3246 (0.9%)	7 (0.2%)	147 (1.3%)	643 (1.3%)	1 (0.1%)
Salinity	1 (0.004%)	17 (0.005%)	-	-	3 (0.01%)	-
Chlorophyll-a	523 (2.0%)	14714 (3.9%)	63 (1.9%)	78 (0.7%)	805 (1.64%)	2 (0.2%)
Fluorescence	10 (0.03%)	273 (0.1%)	1 (0.03%)	-	5 (0.01%)	-
NO_3^-	24 (0.1%)	2166 (0.6%)	4 (0.1%)	13 (0.1%)	213 (0.4%)	-
PO_4^{3-}	398 (1.5%)	19195 (5.1%)	15 (0.5%)	248 (2.2%)	1914 (3.9%)	6 (0.6%)
SiO_2	66 (0.3%)	5878 (1.6%)	-	91 (0.8%)	1303 (2.6%)	1 (0.1%)

Next, we apply EnDED combining the Interaction Information with the Data Processing Inequality method using artificially-generated triplets. Similar to the Jaccard index filter, MICtools networks contained the highest proportion of environmentally-driven associations (9.1-11.4%) followed by SparCC networks (4.3-4.6%), and less environmentally-driven associations in FlashWeave networks (1.0-2.6%). Again, while we detected proportionally more environmentally-driven associations in eukaryotic than prokaryotic networks constructed with MICtools and FlashWeave, it was slightly less in SparCC networks (Table 22 and Figure 28).

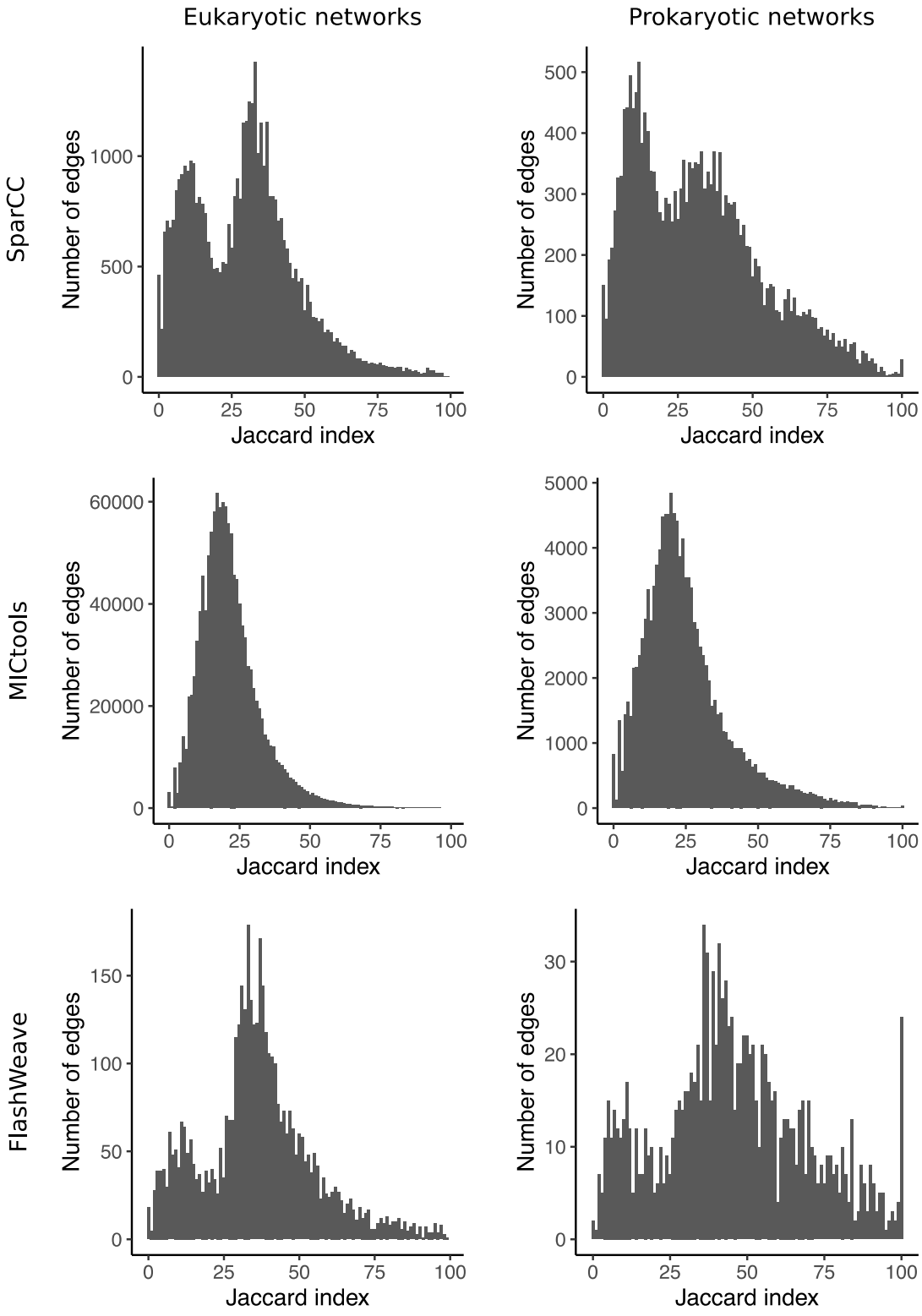


Figure 27**: Jaccard index.** The histograms display the Jaccard index of the associations in the prokaryotic and eukaryotic networks constructed with the SparCC approach implemented in FastSpar, MICtools and FlashWeave.

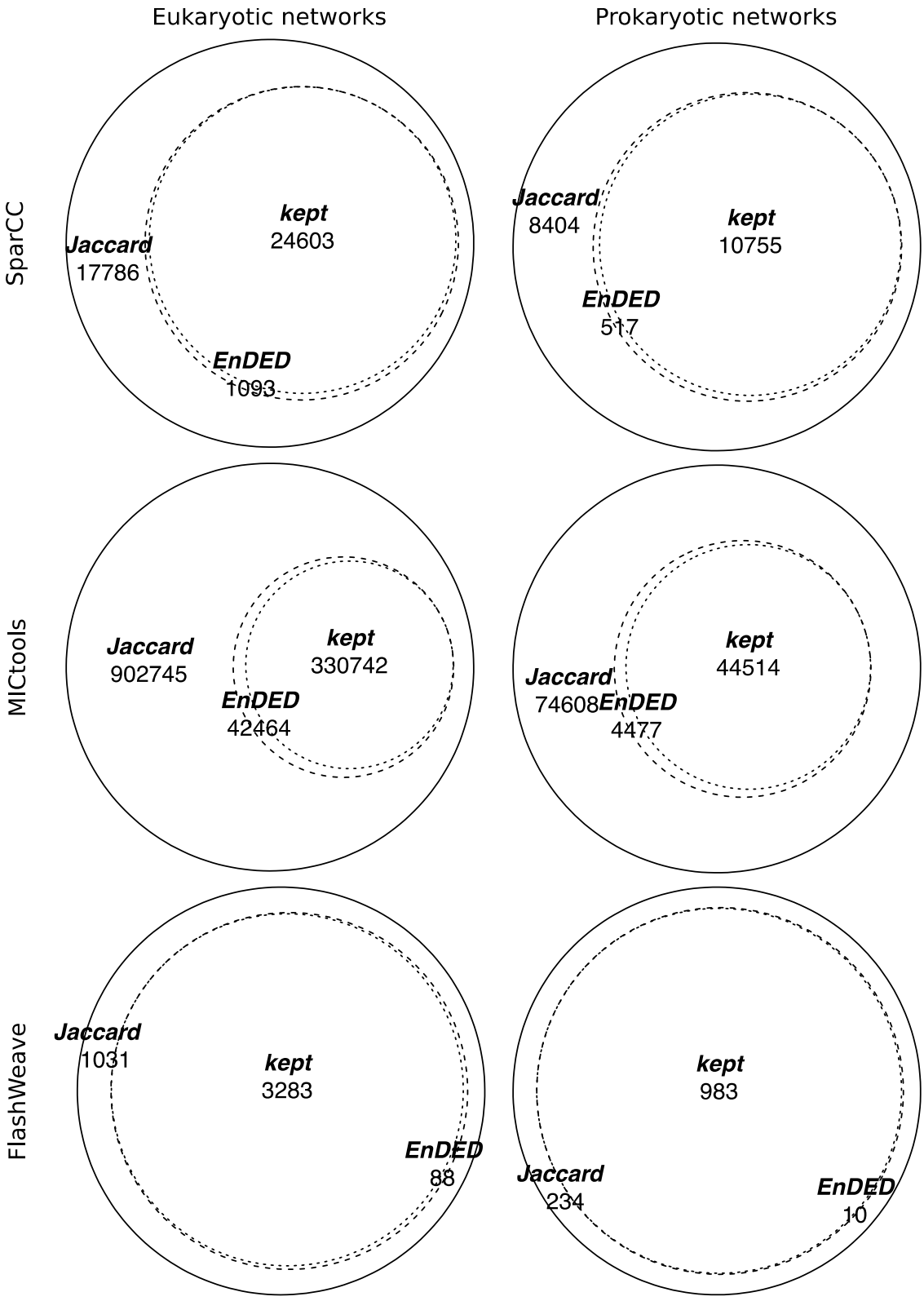


Figure 28**: Kept and removed associations.** We show the number of kept and removed associations in prokaryotic and eukaryotic networks constructed with the SparCC approach implemented in FastSpar, MICtools, and FlashWeave. We used the 25% threshold on the Jaccard index, and removed environmentally-driven associations detected with the tool EnDED. Associations still present in the networks are indicated as kept.

In order to compare eukaryotic and prokaryotic associations, we quantified the number of environmentally-driven indirect dependencies for each environmental factor (Table 22). Surprisingly, Temperature was not the main driver of environmentally-driven associations in the Malaspina Surface networks. However, temperature-driven associations were more prominent among prokaryotic than eukaryotic networks (SparCC and MICtools but not FlashWeave networks). In contrast, we found more total-chlorophyll-a-driven associations in all eukaryotic than prokaryotic networks indicating that the eukaryotic environmental preference causes more indirect dependencies in network inference than prokaryotic environmental preference. The main environmental driver in prokaryotic networks were nutrients as a whole and PO_4^{3-} in particular. The nutrient SiO_2 was responsible for a higher proportion of environmentally-driven associations in the prokaryotic than eukaryotic networks. Nutrients were also main drivers in eukaryotic networks together with the total-chlorophyll-a. However, the results indicate that the network construction tool resulted in different proportions of environmentally-driven associations. Thus, the ranking of environmental factors may vary depending on the tool of choice but generally main patterns emerge. Results also indicate that all tools were prone to generate indirect dependencies, at least to a minor fraction, which should be removed before down-stream analysis and biological interpretation.

Lastly, we determined the number of kept and environmentally-driven associations shared between networks constructed with associations in the prokaryotic and eukaryotic networks (Figure 29). Most associations of the SparCC networks are also included in the MICtools networks. Most kept associations of the FlashWeave networks are included in the SparCC and MICtools networks, but a majority of eukaryotic environmentally-driven associations detected in the FlashWeave network have not been present in the other two networks. Results indicate a greater overlap between the three tools in kept than environmentally-driven associations. The 3014 eukaryotic and 914 prokaryotic common associations have been inferred via each of the three tools and have a Jaccard index above 25% and have not been detected as environmentally-driven. These associations may be promising candidates for potential interactions and further biology-driven investigations. However, in this chapter, we were interested on the number of indirect dependencies among associations that are shared between the three methods in contrast to associations inferred by one or two methods.

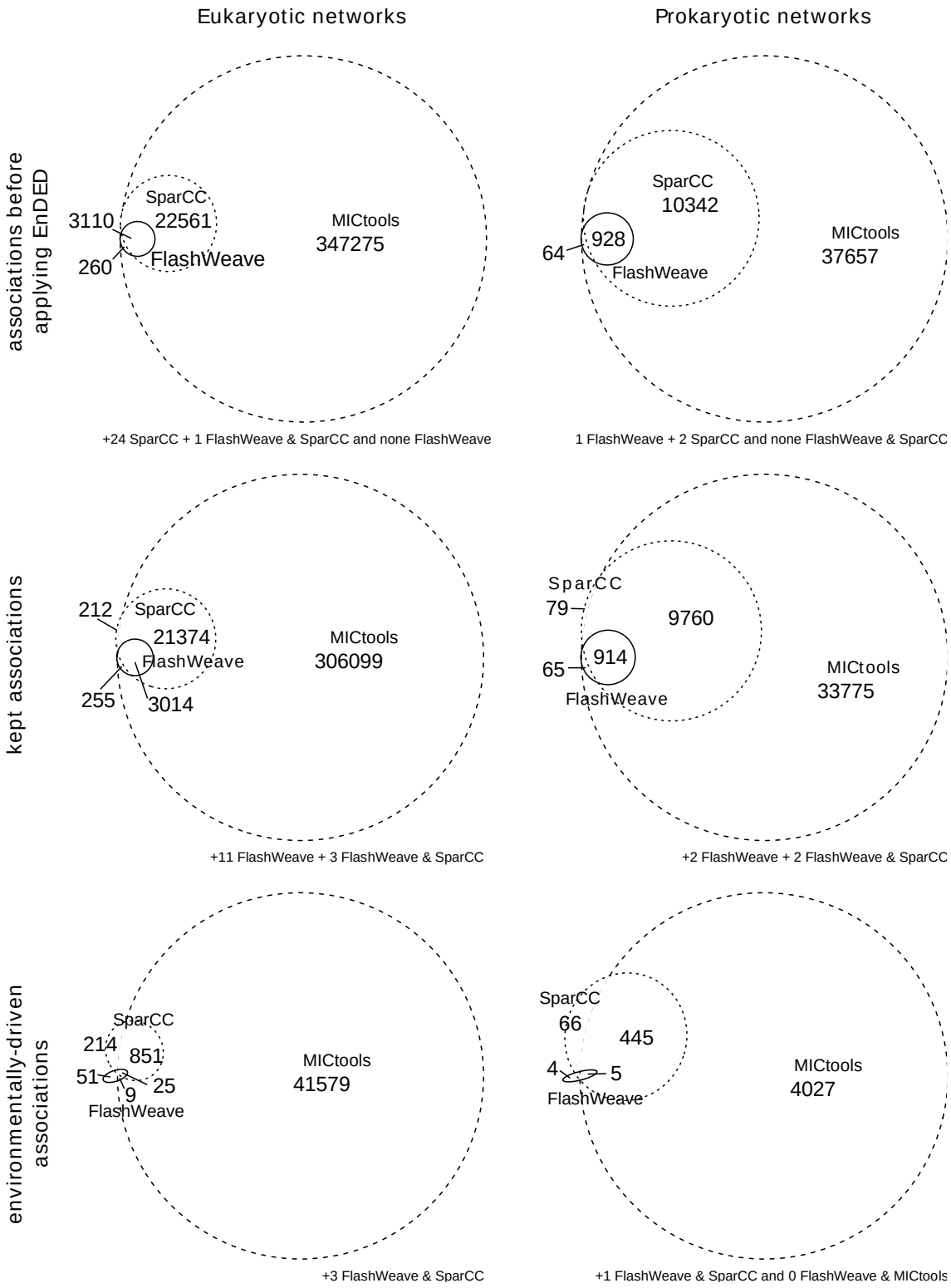


Figure 29**: Shared associations.** Number of shared associations between the eukaryotic and prokaryotic networks constructed with the SparCC approach implemented in FastSpar, MICtools, and FlashWeave. The agreement of edges is shown for networks before applying EnDED, after applying EnDED (kept associations), and for associations detected as environmentally-driven.

We found 3014 kept and 25 environmentally-driven associations appearing in all three eukaryotic networks (120.56x more kept than removed associations). Considering the numbers, the overlap between methods is greater for the kept than the environmentally-driven associations. However, we kept 88.6-97.4% and removed 2.6-11.4% of edges. Thus, we determined whether we also proportionally find more agreement among the kept than removed associations. First, we scaled the numbers by the union of edges that appear in the three networks. The fractions are very low due to the massive number of edges in the MICtools network: 0.91% kept and 0.06% environmentally-driven associations lay in the intersection of methods (15.17x more). We can also scale based on the number of edges that appeared in at least two networks: 12.23% kept and 2.82% environmentally-driven associations lay in the intersection of methods (4.34x more). Lastly, considering the number of edges in the smallest network (FlashWeave network): 91.81% kept and 28.41% environmentally-driven associations appearing in the FlashWeave network also appear in the other two networks (3.23x more). The corresponding values when scaling by the number of associations in the SparCC and MICtools networks, and results for the prokaryotic networks are listed in Table 23. In all scaling procedures, the proportion of associations in the intersection of methods is higher among kept associations.

We showed how EnDED performed on the tools that we have previously used on Malaspina Surface data but other tools could have been included. Here, we determined how EnDED performs on associations detected via three different methods. Our results indicated that all tools were prone to indirect dependencies, at least to a minor fraction. We found less indirect dependencies among associations inferred by all three tools than associations inferred by one or two methods. Future systematic network construction tool benchmarking may include the quantification of environmentally-driven edges for which EnDED provides several methods.

Table 23: **Kept and environmentally-driven associations appearing in the intersection of networks constructed by the three methods.** If the number of shared associations are used, i.e., no scaling, the agreement between methods is 121-183x larger between kept than environmentally-driven associations (highlighted in gray). The discrepancy between them varies depending on the scaling factor. The union of all scales by the number of associations detected it at least one of the three networks. Union of 2-3 indicates that only those edges are considered that are present in at least 2 networks. Lastly, we scaled by the number of edges present in the single networks. In each case, we found a higher fraction of edges in the intersection of kept than environmentally-driven associations.

Scaling	No scaling	Union of all	Union of 2-3	FlashWeave	SparCC	MICtools
Eukaryotes						
removed	25	0.06%	2.82%	28.41%	2.29%	0.06%
kept	3014	0.91%	12.23%	91.81%	12.25%	0.91%
factor	120.56x	15.17x	4.34x	3.23x	5.35x	15.17x
Prokaryotes						
removed	5	0.11%	1.11%	50.00%	0.97%	0.11%
kept	914	2.05%	8.51%	92.98%	8.50%	2.05%
factor	182.8x	18.64x	7.67x	1.86x	8.76x	18.64x

Factors leading to indirect dependencies

> *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.*
> Chapter 5: Disentangling environmental effects in microbial association networks

Ai *et al.* (2019) used data from the SPOT time-series (monthly sampling from 5 m depth from August 2000 to January 20211) including bacterial OTUs (operational taxonomic units), environmental data, and the total bacterial and viral abundance measured by microscopy. Each component (OTUs, environmental factor, bacterial and viral abundance) may act as a mediator for a putative bacterial interaction (so-called three-way associations). The study found that the total bacterial abundance appears to predict correlations of seven pairs of bacterial OTUs showing strongest correlations when the total bacterial abundance is high or low (Ai *et al.*, 2019). The numbers for other (selected) mediating factors are: silicate (12 connections), viral abundance (5 connections), salinity (6 connections), bacterial productivity as measured by thymidine incorporation (4 connections), the rate of change in day length (spring vs fall) (4 connections), phosphate concentration (3 connections), chlorophyll-A concentration (3 connections), and less connections for other environmental parameters such as temperature (2 connections) and nitrate (2 connections) (Ai *et al.*, 2019). Moreover, the study found that four OTUs from the genus *Flavobacteria* appear in most three-way associations (Ai *et al.*, 2019). Given its importance, the work also focused on three-way associations including nodes from the *SAR11* clade (*Pelagibacterales*), which is abundant and exerts many associations. Only one node appears to be involved in several three-way interactions suggesting that SAR11 clades are not active mediators nor governed by mediators in the studied bacterial community (Ai *et al.*, 2019).

Identifying mediators (Ai *et al.*, 2019) is conceptually different from identifying possible drivers of indirect dependencies (Chapter 5). A mediator mediates whether or not an interaction occurs, it allows inferring potential interactions between three components, i.e., going beyond pair-wise associations (between two components). Thus, it is biologically oriented. In contrast, indirect dependencies are errors in the association study and rather mathematically oriented. Nevertheless, the mediator study (Ai *et al.*, 2019) further motivates including parameters such as bacterial abundance or specific microorganisms in environmentally-driven edge detection. Thus, in other works, we included cell-count data (Krabberød *et al.*, 2021) and phytoplankton taxa (Arandia-Gorostidi *et al.*, in preparation) to detect indirect dependencies.

Indirect dependencies detected using cell-count data

"*Long-term patterns of an interconnected core marine microbiota*"

(Krabberød *et al.*, 2021)

Marine microbiotas include core taxa that are usually key for ecosystem function. However, despite their importance, core marine microorganisms are relatively unknown. Most core microbiotas have been defined based on species occurrence and abundance. Yet, species interactions are also important to identify core microorganisms, as communities include interacting species. The work investigates interconnected bacteria and small eukaryotes putatively composing the core microbiota populating the model long-term marine-coastal observatory at the Blanes Bay in the Mediterranean Sea over ten years.

In order to determine interconnected core marine microbiota, several steps have been done. First, we inferred the initial BBMO association network using the tool eLSA (Xia *et al.*, 2011, 2013) as described in previous chapters using the sequence abundance table and 15 contextual abiotic and biotic variables. These variables include the ten environmental factors, which we have used previously,

- Daylength (hours of light);
- Temperature (°C);
- Turbidity (estimated as Secchi disk depth [m]);
- Salinity;
- Total Chlorophyll a [Chla] (µg/l);
- inorganic nutrients (µM): PO_4^{3-}, NH_4^+, NO_2^-, NO_3^-, and SiO_2.

In addition, the environmental factors were extended by five microbial groups via cell-count data,

- heterotrophic prokaryotes [HP] (cells/ml);
- *Synechococcus* (cells/ml);
- total photosynthetic nanoflagellates [PNF; 2-5µm size] (cells/ml);
- small PNF (2µm; cells/ml);
- heterotrophic nanoflagellates [HNF; 2-5µm size] (cells/ml).

Cell counts were done by flow cytometry (heterotrophic prokaryotes, *Synechococcus*) or epifluorescence microscopy (PNF, small PNF and HNF).

Second, a filtering strategy was applied to infer core interactions from the initial network. The filtering strategy combined different filters. To infer environmentally-driven associations, EnDED was run in default mode using network-based environmental triplets appearing in the initial network and the combination of the four methods: Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality. If the four methods agreed that an association was environmentally-driven, then it was removed from the network. In addition, only edges

representing the strongest associations were considered. An association was considered as strong, if

- the absolute value of the local similarity score (LS) is above 0.7;
- the absolute value of the Spearman correlation is over 0.7;
- the significance was below 0.001 (for both, the p-value, and the Bonferroni false discovery rate (q-value);

Lastly, only microorganism present in over 30% of samples were considered. The retained associations were referred to as core associations. ASVs participating in core associations were defined as core ASVs. However, the ASVs involvement in ecological interactions need further experimental validation.

Both core associations and core ASVs constitute the core network, which represents the core microbiota. The core network contained 262 nodes and 1411 edges. It includes only the strongest microbial associations that are inferred during a decade and, according to the set definition, determines the core microbiota. The associations in the core microbiota may represent proxies for species interactions since steps have been taken to remove associations that are driven by environmental factors. This was the aim of EnDED. Thus, EnDED was an important aspect of the filtering strategy in this investigation.

However, in context of this thesis, we want to investigate the usefulness of employing cell-counts in addition to the environmental factors considered in Chapter 5 and Chapter 6. For these previous analyses, we considered microorganisms present in over 15% of samples. Thus, for sake of comparison, here we rerun the analysis using the preliminary BBMO network as used in Chapter 5 and Chapter 6. The network contained 29820 edges. We run EnDED using artificially-generated triplets to detect indirect dependencies. The number of environmentally-driven edges due environmental factors and cell-count data are listed in Table 24.

Results show that the cell-counts detected 1098 indirect dependencies that would have been missed otherwise, i.e., including cell-counts as biotic environmental factors removed 4413 (14.80%) edges in contrast to 3315 (11.12%) when cell-counts are not considered. More indirect dependencies were due to *Synechococcus* and nanoflagellates than nutrients (Table 24). In general, we find a higher fraction of negative indirect associations (Table 24). However, we detected more positive (53.3%) than negative (46.7%) indirect edges due to the number of photosynthetic nanoflagellates cells. Accounting for the type of association partners (bacteria versus eukaryote), more bacterial associations were negative in contrast to more positive associations between eukaryotes and between the two domains (Table 25). These numbers are in contrast to the general trend observed with other cell-count data and environmental factors. For instance, the number of heterotrophic nanoflagellates cells removed mainly negative associations (94.4%). The discrepancy in results considering photosynthetic and heterotrophic nanoflagellates cells, may be because photosynthetic nanoflagellates are much more influenced by environmental

fluctuations than heterotrophic nanoflagellates. Sunlight is one of the main sources of energy of photosynthetic nanoflagellates. Thus, they follow more closely seasonal patterns, especially of day-length and temperature, than the heterotrophic nanoflagellates. Indeed, temperature and day length accounted for 32.5-37.8% of positive indirect dependencies in contrast to other factors (3.7-21.1%).

Table 24: **Number and fraction of environmentally-driven edges for each available environmental factor including cell-counts.** We detected environmentally-driven edges with EnDED using artificially generated environmental triplets, i.e., we introduced artificial edges to connect each ASV with each environmental factor. The network contained 29820 edges: 24458 (82%) positive and 5362 (18%) negative. Using the ten environmental factors removes 4.3% of the positive and 42.2% of the negative edges. Extending the indirect-dependencies detection through cell-counts, removed 6.7% of the positive and 51.8% of the negative edges.

Environmental factor	Indirect	Positive	Negative
10 factors (without cell-counts)	3315 (11.12%)	1053 (31.8%)	2262 (68.2%)
Temperature	1920 (6.44%)	725 (37.8%)	1195 (62.2%)
Total chlorophyll-a concentration	838 (2.81%)	82 (9.8%)	756 (90.2%)
Day length	730 (2.45%)	237 (32.5%)	493 (67.5%)
NO_2^-	192 (0.64%)	26 (13.5%)	166 (86.5%)
SiO_2	162 (0.54%)	6 (3.7%)	156 (96.3%)
NO_3^-	57 (0.19%)	12 (21.1%)	45 (78.9%)
Turbidity	47 (0.16%)	-	47 (100%)
NH_4^+	-	-	-
PO_4^{3-}	-	-	-
Salinity	-	-	-
heterotrophic prokaryotes [HP] (cells/ml)	28 (0.09%)	-	28 (100%)
Synechococcus (cells/ml)	368 (1.23%)	39 (10.6%)	329 (89.4%)
total photosynthetic nanoflagellates [PNF; 2-5µm size] (cells/ml)	979 (3.28%)	522 (53.3%)	457 (46.7%)
total small photosynthetic nanoflagellates [PNF; 2µm size] (cells/ml)	342 (1.15%)	160 (46.8%)	182 (53.2%)
heterotrophic nanoflagellates [HNF; 2-5µm size] (cells/ml)	425 (1.43%)	24 (5.6%)	401 (94.4%)
All 17 factors (including cell-counts)	4413 (14.80%)	1636 (37.1%)	2777 (62.9%)

Table 25: **Photosynthetic nanoflagellates.** Number and fraction of environmentally-driven edges detected through the number of photosynthetic nanoflagellates separated by the type of association partner.

Association partners	Total	Positive	Negative
all	979 (100%)	522 (53.3%)	457 (46.7%)
Bacterial associations	263 (26.9%)	102 (38.9%)	161 (61.2%)
Eukaryotic associations	183 (18.7%)	102 (55.7%)	81 (44.3%)
Bacteria-Eukaryote associations	533 (54.4%)	318 (59.7%)	215 (40.3%)

In conclusion, EnDED integrates well in filter strategies, e.g., when determining core associations. Here, we extended the environmental factors by cell-count data. Cell-count data appears to be a valuable addition to detect environmentally-driven associations, i.e., indirect dependencies among microbial associations that are due to biotic factors. The number of heterotrophic prokaryotes, *Synechococcus*, and heterotrophic nanoflagellates cells infer much more negative than positive edges, similarly to other environmental factors. In contrast, the number of photosynthetic nanoflagellates cells remove a similar fraction of positive and negative

associations. In sum, cell-count data are a valuable extension to the detection of environmentally-driven edges, especially positive associations.

Indirect dependencies between bacterioplankton due to phytoplankton

"*Biotic interactions between phytoplankton and heterotrophic bacteria dominate microbial seasonal dynamic in coastal ocean waters*"

(Arandia-Gorostidi *et al.*, in preparation)

Marine phytoplankton represents one of the most important carbon sources for another microbial group, the heterotrophic bacterioplankton. Therefore, the interactions between them may largely determine biogeochemical cycling in the coastal ocean. However, how these interactions impact their temporal dynamics is still unclear. This work employs network analysis of a three-year time-series dataset to assess associations between different phytoplankton and bacterial taxa. Expanding our view about bacteria-phytoplankton interactions, this work shows that the specific composition of co-occurring phytoplankton may largely mediate patterns of bacterial seasonal reoccurrence.

The time series contained 42 samples (monthly sampling from July 2009 to December 2012). Here, we consider bacterial OTUs (0.2-3µm), phytoplankton taxa (determined via microscopy), and the environmental factors: temperature, salinity, and fluorescence. The network was constructed with the local similarity analysis (LSA) program (Ruan *et al.*, 2006) and contained 2062 edges (1976 microbial associations and 86 from a microorganism to an environmental factor).

Before applying EnDED to detect indirect dependencies, rare microorganisms were removed keeping 11 phytoplankton taxa and 152 bacterial OTUs. EnDED was applied twice using the network-based triplet approach combining the methods Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality. First, EnDED was used to detect indirect dependencies among the 1976 microbial associations, i.e., between microorganisms including phytoplankton and heterotrophic bacterioplankton, due to three environmental factors. We found 943 microbial associations appearing in at least one network-based environmental triplet, in total 966 triplets, and 127 indirect edges according to the combination approach using all four methods. Using a single method would infer more indirect edges (data not shown).

Second, EnDED was used to detect indirect dependencies between the 1911 bacterial associations, i.e., associations between heterotrophic bacterioplankton, due to the 11 phytoplankton taxa. We found 137 environmental triplets including seven of the considered eleven phytoplankton taxa. Several edges have been detected as environmentally-driven according to single methods (data not shown) but only 3 edges according to the combination approach using all four methods. The method restricting the most was Data Processing Inequality.

The number of triplets and environmentally-driven edges are listed in Table 26. Temperature was the top driver accounting for all 127 environmentally-driven associations, 38 (29.9%) positive and 89 (70.1%) negative associations. Two of them were between a bacterial OTU and the phytoplankton taxon number 264. The very same phytoplankton taxon was the only detected driver of three environmentally-driven bacterial negative associations in the second run of EnDED. All three edges were also in a triplet with temperature, but only one of them was also environmentally-driven according to temperature (OTU0117 and OTU0005), the other two only due to the phytoplankton taxon (OTU0016 and OTU0014, and OTU0016 and OTU0033).

Table 26: **Number and fraction of environmentally-driven edges for each available environmental factor including phytoplankton taxa.** We detected environmentally-driven edges with EnDED using network-based environmental triplets, i.e., environmental factors have been included in the network construction.

Environmental factor	Network-based environmental triplets	Environmentally-driven edges
Temperature	928	127 (38 positive and 89 negative)
Salinity	37	0
Fluorescence	1	0
Phy_sp_035	6	0
Phy_sp_036	51	0
Phy_sp_178	27	0
Phy_sp_185	6	0
Phy_sp_195	-	-
Phy_sp_196	-	-
Phy_sp_221	-	-
Phy_sp_243	-	-
Phy_sp_244	3	0
Phy_sp_254	10	0
Phy_sp_264	34	3 (3 negative)

The aim of EnDED in (Arandia-Gorostidi *et al.*, in preparation) was to remove environmentally-driven edges due to environmental factors and phytoplankton taxa. However, in context of this thesis, using EnDED with microbial triplets (two bacterial OTUs and one phytoplankton taxon) was a first step and test if general microbial triplets, as suggested in (Deutschmann *et al.*, 2020) should be considered in future studies. In sum, EnDED detected 129 (6.5%) indirect edges among the 1976 microbial edges. Temperature was the main driver as it was in the BBMO network investigations employing network-based triplets (Deutschmann *et al.*, 2020; Krabberød *et al.*, 2021) and artificially-generated triplets (Deutschmann *et al.*, 2021).

Here nutrients were not considered although the dataset contained inorganic nutrient concentration (nitrate and phosphate) and photosynthetically active radiation (PAR). Future investigations may find more environmentally-driven edges using artificially-generated triplets and incorporating other environmental factors such as nutrients. Using phytoplankton to detect indirect dependencies was not fruitful, promising at best. Only 3 (2.2%) of 137 triplets were inferred as indirect in contrast to 127 (13.1%) of 966 triplets with abiotic factors. However,

considering only the phytoplankton taxon number 264, we inferred 3 (8.8%) of 34 triplets as indirect in contrast to 127 (13.7%) of 928 triplets with temperature.

To conclude, the results indicate, that it is still worth considering general microbial triplets in future studies. While we investigated two types of microorganisms (phytoplankton affecting bacterial associations), future investigations should also consider triplets of the same type, e.g., three connected bacterial OTUs.

Final remarks

- ⇒ EnDED allows indirect dependency detection using network-based and artificially-generated triplets.
- ⇒ We quantified more environmentally-driven edges when using artificially-generated triplets instead of network-based environmental triplets.
- ⇒ Comparing both approaches, we detected similar numbers of environmentally-driven associations for temperature and day-length but more associations for nutrients and the total chlorophyll-a concentration using artificially-generated triplets.
- ⇒ Applying EnDED to networks constructed with different tools indicated that all tools were prone to indirect dependencies, at least to a minor fraction.
- ⇒ We found less indirect dependencies among associations inferred by three tools (SparCC, MICtools, and FlashWeave) than associations inferred by one or two methods.
- ⇒ Future systematic network construction tool benchmarking may include the quantification of environmentally-driven edges for which EnDED provides several methods.
- ⇒ EnDED integrates well in filter strategies, e.g., when determining core associations or potential interaction partners.
- ⇒ Extending the environmental factors by cell-count data appears to be a valuable addition to detect indirect dependencies among microbial associations that are due to biotic factors.
- ⇒ The number of heterotrophic prokaryotes, *Synechococcus*, and heterotrophic nanoflagellates cells infer much more negative than positive edges, similarly to other environmental factors.
- ⇒ The number of photosynthetic nanoflagellates cells remove a similar fraction of positive and negative associations.
- ⇒ Investigating two types of microorganisms (phytoplankton affecting bacterial associations) revealed overall little indirect dependencies.
- ⇒ Nevertheless, including microbial triplets remains promising and future investigations should consider triplets of the same type, e.g., three bacterial associations.

Part III Further discussion and thesis conclusions

In this section, we extend the previous discussion. Then, we present and discuss future perspectives and directions. Finally, we conclude the thesis.

Chapter 9 Environmentally-driven associations

Negative versus positive environmentally-driven associations

We applied EnDED on networks using different tools and data. For instance, for temporal data, we applied (extended) local similarity analysis, i.e., LSA and eLSA, (Ruan *et al.*, 2006; Xia *et al.*, 2011, 2013) to construct

- the simulated networks;
- the BBMO network;
- the phytoplankton-bacterioplankton network.

Further, for spatial data, we applied different tools

- the global network (compilation of Malaspina and Hotmix data) and Malaspina Surface network constructed with FlashWeave (Tackmann *et al.*, 2019);
- the Malaspina Surface network constructed with the method SparCC (Friedman & Alm, 2012) as implemented in FastSpar (Watts *et al.*, 2019);
- the Malaspina Surface network constructed with the tool MICtools (Albanese *et al.*, 2018).

In all analysis, we found a higher fraction of negative than positive environmentally-driven edges except when considering the number of photosynthetic nanoflagellates cells on the BBMO network. Associations may represent positive or negative interactions, but they can also indicate high niche overlap (positive association) or divergent niches (negative association) between microorganisms (Hernandez *et al.*, 2021). We hypothesize that most of the removed negative edges represented associations between microorganisms from divergent niches, i.e., several negative associations are probably due to different environmental preferences (different niches) of microorganisms.

In Chapter 5, we showed that the Jaccard index representing a level of microbial co-occurrence scored equal or below 50% for most negative associations. In Chapter 6 and Chapter 7, we included the Jaccard index in our filtering strategy and removed proportionally more negative than positive edges. For example, our preliminary BBMO network (significant associations derived with eLSA) contained 18% negative edges compared to 0.9% in the single static BBMO network (after applying EnDED and Jaccard index). Furthermore, we tested the Jaccard index-based approach in the global network study generating sample-specific subnetworks. It was tested robust. However, since the approach uses simple microbial presence and absence (sequence abundance greater or equal to zero), possible future evaluations and technical investigations may introduce a cut-off level instead of zero.

Conclusion

A higher fraction of negative than positive associations was found to be environmentally-driven. Negative association may be true negative interactions or represent contrary environmental preference. Our results indicate that most negative associations in association networks may rather represent the latter, i.e., environmental preferences rather than true interactions.

Factors causing indirect microbial associations

For the temporal BBMO data and the spatial dataset compilation, we used available environmental factors, but not all factors that could affect microbial dynamics and potentially could generate indirect edges. For instance, indirect edges associated with biotic interactions (e.g., two bacteria sharing a positive edge as they are symbionts in the same protists) were not considered. Applying EnDED in a filtering strategy using cell-counts (Krabberød *et al.*, 2021) and phytoplankton taxa (Arandia-Gorostidi *et al*., in preparation), detected indirect edges that may be due to biotic influences. In addition, larger single-celled eukaryotes should be considered. For instance, we excluded larger single-celled eukaryotes by filtering the BBMO water at 20 µm, but there is a substantial diversity of larger protists (de Vargas *et al.*, 2015) with potential interactions among them and to smaller-sized microorganisms (Lima-Mendez *et al.*, 2015). The literature-based protist interaction database PIDA includes many groups that are generally larger than the filters used in BBMO, e.g., *Radiolarians* and many *Dinoflagellates* (Bjorbækmo *et al.*, 2019). Moreover, viruses are ubiquitous in marine environments (Endo *et al.*, 2020). Viruses infect diverse eukaryotes and such interactions may be important in biogeochemical processes in the ocean (Endo *et al.*, 2020). Viruses have been included in network-based microbial investigations (Chow *et al.*, 2014; Needham *et al.*, 2017), a number of interactions may be mediated either by the viruses themselves or an unmeasured factor that relates to viral abundance as detected via extended liquid association (ELA) analysis (Ai *et al.*, 2019).

Conclusion

Various factors may cause indirect dependencies in association networks such as abiotic factors (e.g., temperature), nutrients, but also biotic factors (e.g., specific microorganisms, other living organisms or viruses). Thus, on one hand, future sampling should expand metadata collection in order to account for (more) abiotic and biotic factors that could explain and identify indirect dependencies. On the other hand, an updated version of EnDED should allow to investigate any triplet, e.g., microbial triplets.

Technical aspects of environmentally-driven associations

We found more environmentally-driven associations in our global network spanning the water column from the surface to the deep ocean as opposed to a global network covering the epipelagic (surface and DCM layer) (Lima-Mendez *et al.*, 2015). However, such comparison is unfair because we sampled mainly picoplankton from the surface to the deep ocean while (Lima-Mendez *et al.*, 2015) samples cover the epipelagic zone including several size fractions. Compared to the BBMO network based on ten years of data, our global ocean network showed a lower fraction of environmentally-driven edges. The BBMO environmental table was more complete in contrast to several missing environmental datapoints in the global ocean dataset. We list the fraction of environmentally-driven associations in these three networks (Table 27), but refrain from further comparisons because each of the three studies used different sampling, environmental data, filter and network construction strategies.

Table 27**: Fraction of environmentally-driven edges** for environmental factors in this and previous works.

	Global ocean	**Global Ocean**	**Time series**
Samples	Epi- (surface and DCM), meso-, and bathypelagic	Epipelagic (surface and DCM)	Coastal surface
Campaigns or sites	Malaspina & Hotmix	Tara Oceans	Blanes Bay Microbial Observatory
Considered edges	All edges with all environmental factors	Edges in environmental triplets	All edges with all environmental factors
Method			
Combination of Interaction Information and Data Processing Inequality	8.9%	-	11.1%
Interaction Information	53.7%	12% (37% of triplets)	86.4%
Environmental factor / Method	Combination	Interaction Information	Combination
Fluorescence/Chlorophyll-a concentration	Top 6 (0.01%)	-	Top 2 (2.8%)
Day length	-	-	Top 3 (2.5%)
NH_4^+	-	-	0
NO_2^-	-	Top3	Top 4 (0.6%)
NO_3^-	Top 1 (4.9%)	-	Top 6 (0.2%)
PO_4^{3-}	Top 2 (4.2%)	Top 1	0
SiO_2	Top 3 (2.0%)	Top 6	Top 5 (0.5%)
Salinity	Top 5 (0.2%)	-	0
Temperature	Top 4 (1.9%)	Top2	Top 1 (6.4%)
Turbidity	-	-	Top 7 (0.2%)
Reference	Chapter 7	(Lima-Mendez et al., 2015)	Chapter 6

Using the Malaspina surface data, we constructed eukaryotic and prokaryotic networks with three different tools (Chapter 8). We found all tools were prone to effects of indirect dependencies, at least to a minor fraction. However, depending on the tool, we found more environmentally-driven associations among eukaryotic or prokaryotic associations, and the

ranking of top environmental-drivers differed. Our results indicate that different tools result in different levels of indirect dependencies. There remains an open question of how, how much, and which technical and biological factors cause indirect dependencies in the dataset.

Conclusion

The effect of environmentally-driven edges is prominent in association networks. All tested network construction tools were prone to effects of indirect dependencies but the extent of specific environmental drivers differed among tools. Thus, comparing environmental-drivers in microbial ecosystems will require same sampling, environmental data, filter and network construction strategies.

Chapter 10 Analyzing networks

Usage of environmental data in network analysis

Microbial community composition is strongly influenced by multiple environmental factors. The prime example is temperature that exerts selection on the ocean microbiome (Sunagawa *et al.*, 2015; Ibarbalz *et al.*, 2019; Salazar *et al.*, 2019; Logares *et al.*, 2020). In microbial association networks, an edge may not represent a true interaction but a common or opposite response to environmental factors. Thus, network-based microbial investigations should account for environmental influence. There are different strategies on how to use environmental data in network analysis prior-, during-, and post-network construction.

Several prior-network construction methods aim to account for environmental influence. One strategy splits samples into groups and constructs networks for each sample group. For example, a previous work (Mandakovic *et al.*, 2018) constructed two networks representing bacterial soil communities from two areas displaying different pH, temperature, and humidity gradients. 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. A temporal network-based investigation (Lambert *et al.*, 2021) constructed networks for three consecutive years including an average year (physical and chemical parameters were close to the long term mean), and two years that were marked by environmental perturbations. However, defining the groups may not be straight forward, and splitting samples may leave too few samples for network construction. Another pre-network construction approach aims to regress out the influence of environmental factors on microbial abundances by using the residual microbial abundances to infer associations (Warton *et al.*, 2015). However, Faust (2021) argues that many species respond nonlinearly to environmental parameters and extending regression to handle nonlinearities may increase the risk of overfitting the data.

During network construction, environmental components may be included as nodes in the networks among microbial nodes. Several network construction tools allow to include environmental factors, e.g., LSA and eLSA (Ruan *et al.*, 2006; Xia *et al.*, 2011, 2013), CoNet (Faust & Raes, 2016), MICtools (Albanese *et al.*, 2018), and FlashWeave (Tackmann *et al.*, 2019). Previous marine microbial association studies found more microbial associations than associations between microorganisms and environmental factors (Steele *et al.*, 2011; Lima-Mendez *et al.*, 2015; Krabberød *et al.*, 2021), which indicates the dominance of microbial associations over associations between microorganisms and environmental factors.

Post-network construction, environmental nodes may be used to filter indirect associations via environmental triplets (Lima-Mendez *et al.*, 2015; Deutschmann *et al.*, 2020). However, introducing environmental factors as nodes in the network has also its flaws. We detected more environmental effects using artificially-generated compared to network-based

triplets. That is, using network-based environmental triplets potentially misses the detection of indirect dependencies among microbial associations indicating that artificially-generated triplets could be a better option. During this thesis, we used several post-network construction approaches to combine environmental information with microbial associations: i) employing environmental data to detect environmentally-driven associations; ii) correlating environmental factors with global network metrics; and iii) investigating if similar network topologies also align with similar environmental values. Another post-network construction approach studied environmental influence by designing a network attack procedure combining environmental tolerance range inference with network stability analyses (Chaffron *et al.*, 2020). It mimics the potential effect of each environmental parameter's variations onto the network. It is used to simulate the effects of environmental changes and to predict their impact on the stability of plankton community structures. Finally, environmental factors can be correlated with detected microbial modules as implemented in the R package WGCNA (Langfelder & Horvath, 2008) and, e.g., used to relate modules to carbon export in the global ocean (Guidi *et al.*, 2016). Other works use environmental data to characterize specific nodes, e.g., to determine microorganisms' environmental preferences (Krabberød *et al.*, 2021; Lambert *et al.*, 2021; Latorre *et al.*, 2021).

Conclusion

Incorporating environmental data as additional nodes during network construction may miss important associations between environmental factors and microorganisms. Instead, it may be valuable to use different approaches to incorporate environmental data, e.g., inferring environmentally-driven associations via artificially-generated triplets and elucidating environmental influence on network architecture by relating environmental factors with network properties.

Drivers of network architecture

Röttjers and Faust (2018) claimed that the effects of environmental variables on network metrics are unclear. A bi-weekly sampling during winter months allowed examining three winter networks of subsequent years, which revealed weather perturbations in temperature or salinity may have altered network topology (Lambert *et al.*, 2021). In Chapter 6, we introduced a monthly subnetwork approach that allows to identify potential environmental drivers of network architecture. As discussed in Chapter 6, correlation analysis pointed to common factors known to influence microbial abundance: temperature and day length (Bunse & Pinhassi, 2017; Giner *et al.*, 2019; Lambert *et al.*, 2019), and to a lesser extent inorganic nutrients (Estrada, 1996; Sala *et al.*, 2002). Furthermore, in Chapter 7, we followed another approach by mapping environmental factors onto the minimal spanning tree (MST) showing main patterns between similar sample-specific subnetworks. The tendency of similar environmental variables to locate together in the

MST indicates a possible connection between environment and network topology since the MST was constructed using network similarity based solely on network topology. Thus, our analyses are a step forward to elucidate the effects of environmental variables on network metrics and network topology; although, we did not consider several other variables that could affect microbial communities and network architecture (as already discussed in previous chapters).

Conclusion

Sample-specific (including monthly) subnetworks provide a step forward towards disentangling the effects of environmental variables on network metrics and network architecture.

Quantifying microbial associations expands their characterization

Microbial associations can be assigned properties, usually significance and association strength. However, these may not be sufficient to select promising associations for interaction hypotheses because increasing significance still yields numerous associations and an association strength cut-off was shown to be not sufficient to separate true from false interactions in simulated data in Chapter 5. Recurring associations may yield potential interaction hypotheses, e.g., when detected via different network construction tools. Weiss *et al.* found little overlap between different tools and methods (Weiss *et al.*, 2016). Thus, associations that are repeatably captured (by different tools/methods on the same dataset or networks constructed for different datasets) may indicate a strong biological signal. Using EnDED (Chapter 8) on three eukaryotic and three prokaryotic networks constructed with three different tools, showed that edges detected as environmentally-driven overlap less than edges that are not environmentally-driven and subsequently, promising potential microbial interactions.

This thesis expands the tool set for characterizing associations to quantifying temporal and spatial recurrence by using sample-specific subnetworks to disentangle association networks aggregated over time and space, respectively. Furthermore, monthly recurrence and region-depth specific recurrence may allow detecting associations prevalent at a certain time and region, respectively, which may provide another approach to infer essential associations. Such essential marine microbial association may represent the core network of the global ocean.

Identifying core networks is not straightforward as discussed in the ninth challenge of a recent perspective (Faust, 2021): First, a microbial network should be constructed for each sample group representing a location, environmental condition, or time point. Second, the intersection of these networks should be computed. However, such a core network is only informative if it has more edges than expected by chance but it is unclear which null model to choose to compute the random expectation (Faust, 2021). Moreover, a global intersection approach discards associations that are more frequently encountered in a subset of specific networks than expected by chance

(Faust, 2021). Our approach of quantifying microbial associations circumvents the all-or nothing global intersection approach but future developments should (include a strategy to) determine if the (temporal or spatial) recurrence is higher or lower than expected by chance.

Conclusion

Focusing on the most significant and strongest associations may not be sufficient to select candidate interactions. Quantifying recurring associations may be promising and several approaches can be done: i) recurring associations using different network construction tools, ii) recurring associations using different datasets, iii) temporal recurrence, and iv) spatial recurrence. Quantifying microbial association recurrence may be a step towards identifying a core network, i.e., associations that are preserved across spatial and temporal scales.

Temporal and spatial patterns

Considering global network metrics, different results indicate that microbial communities are more clustered (higher transitivity) in colder waters compared to warmer counterparts when considering i) samples along the entire water column (Chapter 7), ii) horizontally distributed samples from pole to pole (Chaffron *et al.*, 2020), and iii) monthly data over ten years (Chapter 6). These results suggest that either microorganisms interact more in colder environments or that their recurrence is higher due to a higher environmental selection (exerted by temperature or other factors) increasing their tendency to co-occur. Alternatively, limited resources (mostly nutrients) may prevent the establishment of several microbial interactions in surface vs. deep stratified ocean waters (Chapter 7) or in warmer summer vs. colder winter in temperate areas (Chapter 6). In accordance, using temporal and spatial recurrence, we found more highly prevalent associations in colder than warmer months (Chapter 6) and in the warmer epipelagic than in colder deeper ocean zones (Chapter 7), respectively. It remains to be tested whether this finding is due to biological or technical reasons.

Associations are more repeatable at colder versus warmer months in the ten year BBMO model marine microbial ecosystem using monthly subnetworks. Moreover, the temporal BBMO network appears to collapse from colder to warmer months and reassemble from warmer to colder months. Yet, these findings should be treated with care and investigated using monthly networks. For example, Lambert *et al.* (2021) sampled the Banyuls Bay microbial observatory (SOLA), a coastal site of the North-Western Mediterranean Sea, twice a week during three winters (January–March), and constructed one network per winter. These networks show clear differences in topology between years and microorganisms changed association partners between the years (Lambert *et al.*, 2021). Such investigation among summer months should be done to investigate if microbial communities interact i) more, ii) similarly, or iii) less in summer or winter.

Using temporal BBMO data (ten years) and monthly subnetworks (Chapter 7), we found that most global network metrics indicated a periodicity of one year changing between colder and warmer waters. This is in accordance with previous work that found yearly recurrence in microbial community composition at the BBMO (Giner *et al.*, 2019; Auladell *et al.*, 2020; Krabberød *et al.*, 2021), and at the Bay of Banyuls (Lambert *et al.*, 2019), which is also in the North-Western Mediterranean Sea. In contrast to our work finding repeatable network properties, the three subsequent years revealed inter-annual variations of network topology (Lambert *et al.*, 2021). However, the three years were selected because they displayed environmental changes for which differing network topology should be expected. The strategies of both studies should be combined, investigating several (ten rather than three) years of season-specific networks to determine the change of network topology.

Using spatial data (compilation of Malaspina Surface, Malaspina Vertical Profiles, Malaspina Deep Ocean, and Hotmix) and sample-specific subnetworks, there was no one general up or down trend in global network metrics along the water column from the warmer surface to the colder deep ocean. However, using spatial recurrence, we identified an increasing number and fraction of regional association from the surface to the deep ocean. In Chapter 7, we found most deep-ocean ASVs already appeared in upper layers, but most deep-ocean associations did not appear in upper layers. Similarly, although ASVs re-occur, their associations do not re-occur in the three winter networks from three consecutive years (Lambert *et al.*, 2021). Precisely, a total of 42 ASVs were common to the three high-frequency winter samplings but all of them changed neighbors between years.

Conclusion

Associations may be more repeatable at colder versus warmer waters, but it remains to be tested using a higher sampling frequency allowing the construction of season-specific networks over several years and depth-specific networks of one location, respectively. Moreover, such season-specific networks are the next step to test if recurring annual patterns in network architecture may be due to similar environmental variables. In accordance, depth-specific patterns and changes (up- and down-ward trends) may be better studied using networks constructed per depth and region instead of using subnetworks derived from a single static network. Yet, our approaches represent the first steps in elucidating similar and different network-based patterns between seasons, depths, and between the ocean and the sea (here Mediterranean Sea).

Chapter 11 Additional technical perspectives

Single networks versus subnetworks

From a technical perspective, our monthly subnetwork approach allowed us to see what the single static BBMO network captured since all our temporal network observations are linked to it. Similarly, the sample-specific subnetwork approach allowed us to see what the single static global ocean network captured since all observations are linked to it. Comparing our networks with previous work, main global network metrics (considering edge density, transitivity, and average path length) fall within the range of published networks (Steele *et al.*, 2011; Chow *et al.*, 2013, 2014; Cram, Xia, *et al.*, 2015; Lima-Mendez *et al.*, 2015; Chaffron *et al.*, 2020). However, it remains to be tested whether a single static network or subnetworks derived from it provide not only a different but also better view. A follow-up investigation (not shown in thesis) comparing modules of the single static BBMO network with modules of the temporal network provided no evidence to such claim and further investigations should be done. However, monthly and sample-specific subnetworks provided a new approach suited to quantify association recurrence and investigate the relationship between environmental factors and network topology.

When we mapped depth-related spatial recurrence onto the single static global network visualization (Figure 18), we identified patterns of specific connected ASVs for certain regions. However, our global ocean analysis comprised a well-sampled surface layer of the global ocean and well-sampled MS layers, while fewer samples were available for deeper layers of the global oceans due to the challenges of sampling this habitat. This unevenness of samples may have introduced biases in our study, e.g., when constructing the global static network or determining sample-specific subnetworks. Another bias may be related to the DNA filtration process, which used different filter types (pore size and diameter of the filter). These biases are expected when combining different datasets from large-scale surveys involving complex logistics and sampling procedures. However, we aimed at mitigating possible biases by matching size fractions among datasets (see Methods in Chapter 7) and by performing different tests, which supported that the main reported patterns are robust.

Furthermore, the single static networks may have missed microbial associations that we are not able to infer from our data, and subsequently, did not appear in subnetworks. In the future, we might elucidate them through extensive sampling, e.g., within one time period or region. Thus, for the BBMO network study we refrained from concluding that more microbial associations characterize winter months but we concluded that we could infer more microbial associations in the colder months than in the warmer months. Our suggestion for a higher sample frequency at the BBMO is in contrast to the results from two decades ago. The sampling at the BBMO started in early 2000 using fingerprinting techniques (Schauer *et al.*, 2003). Schauer *et al.* (2003) compared weekly with monthly samples during late winter periods, which demonstrated that

monthly frequency is sufficient to study the variability of the bacterial assemblage. However, in our study we observed a sharp decrease of associations from late winter to early spring and a finer sampling frequency may help to find in-between configurations or pinpoint the moment when the system collapses.

The combination of a single static network (our 1st condition) and microbial co-occurrence (our 2nd condition) has been used previously (Chow *et al.*, 2013; Chaffron *et al.*, 2020). Chow *et al.* (2013) also use the tool eLSA to construct a single static network from ten-year of data but did not exploit the potential to derive monthly subnetworks. Chaffron *et al.* (2020) also use the tool FlashWeave to construct a single static network from an epipelagic global ocean dataset from pole to pole. Since microbial co-occurrence is a necessary but not sufficient condition for a potential interaction to realize itself (Poisot *et al.*, 2012), an interaction is not automatically exhibited when both microorganisms are present in a sample. Thus, our approach generating monthly subnetworks contained the 3rd condition using association duration (Chapter 6) and the approach to generate sample-specific subnetworks used region and depth specific Jaccard index (Chapter 7). However, the association duration condition operates on a time window with one start and one end point in time and, subsequently, does not consider seasonal on-off switches. In addition, our approaches do not take into account possible switches of associations from negative to positive, e.g., when transitioning from low to high-stress environments (Piccardi *et al.*, 2019; Hernandez *et al.*, 2021). If associations switch from negative to positive, a single static network may miss or infer weak associations (Hernandez *et al.*, 2021). Such biases are reduced with higher sampling frequency allowing the construction of time and location specific networks.

Conclusion

Future studies with higher sampling frequency may be able to construct networks within a month (monthly networks instead of subnetworks) and location (location-specific networks instead of subnetworks). Although our approaches are a good starting point that allow us to move forward, they have limitations suggesting caution when making biological interpretations from the temporal network and sample-specific subnetworks.

The lack of gold-standards

No ideal (gold-standard) network construction method exists and some tools are better suited than others for specificities of a dataset, e.g., temporal vs. spatial and homogenous vs. heterogenous. For the three main projects, we decided to use two construction tools, namely eLSA (Xia *et al.*, 2011, 2013) and FlashWeave (Tackmann *et al.*, 2019), which we selected according to the characteristics of our datasets. The former incorporates the temporal aspect of the BBMO dataset of 10 years. Due to missing datapoints, the data was first rarefied and then the missing abundances

were interpolated. However, microbial data is compositional and future studies may apply clr-transformation instead of rarefication (Gloor *et al.*, 2017). We used FlashWeave for the global ocean dataset. Here we applied the clr-transformation. FlashWeave is ideal for our dataset because of its sparse and heterogenous nature. FlashWeave accounts for zeros in the dataset avoiding spurious associations due to many zeros in common, which we have observed with other tools, e.g., in a MICtools (Albanese *et al.*, 2018) surface network of the global ocean. We used this network to determine potential interaction partners for MAST-4 species (Latorre *et al.*, 2021). We aimed to reduce spurious associations using a post-network construction filtering strategy (Latorre *et al.*, 2021). Furthermore, whereas eLSA determined pairwise associations individually, FlashWeave models the network as a whole. Thus, removing some of the players may interfere with the associations of the remaining ones in network constructed with FlashWeave but not eLSA.

There are several approaches predicting microbial interactions. Li *et al.* (2016) state that integrating multiple approaches (co-occurrence patterns, metabolic reconstruction, and mining the literature) could improve the accuracy of microbial interaction prediction. However, there is currently no gold-standard dataset of interactions, i.e., validated interactions and validated cases of two microorganisms that do not interact with each other. However, a few data sets recording known interactions have been generated. For instance, to evaluate the global epipelagic interactome network, a list of known eukaryotic phytoplankton interactions has been compiled (Lima-Mendez *et al.*, 2015). One effort is the website http://aquasymbio.fr/, which is dedicated to collecting associations in aquatic systems (marine and freshwater). Another effort gathered marine microbial interactions described in the literature, PIDA, the Protist Interaction Database (Bjorbækmo *et al.*, 2019). PIDA contains eukaryotic (protist) interactions and interactions between eukaryotes and prokaryotes but no prokaryotic interactions. The data is not complete and lacking interactions may be due to the interaction not occurring or not being detected and described in the literature. Thus, observed interactions cannot be used to determine the accuracy of network inference tools but only their sensitivity (probability that a known interaction is inferred) (Faust, 2021).

Another promising approach is the use of single cells as evidence for potential interactions. Emerging technologies retrieve genomic and transcriptomic information from individual microorganisms: single-cell genomics and single-cell transcriptomics. There is no need for culturing as the approach can use cells taken directly from natural environments. Single Amplified Genomes (SAGs) contain the genomic information of a single cell. Assuming no contamination of the samples, if DNA from different species appears in the same SAG, it implies a potential interaction between the two different microorganisms. For example, finding the bacterial DNA in a eukaryotic SAG suggests that the latter was a host or predator of the former. Assume the network contains an edge between species A and B. If both association partners are

found within the same cell, we have additional evidence that there may be an ecological interaction, e.g., prey-predator or parasite-host. Also, some ecological interactions may require a tied physical connection that may not be separated during single cell isolation leading to finding two microorganisms within the same single cell. Of course, finding genetic material of two organisms within the same single cell is not the idea and purpose of single cell genomics. Instead of ecological interaction the reason behind finding additional organisms may be contamination. In addition, the obtained genetic material is not complete. Thus, although the idea seemed simple, the execution is challenging. However, SAGs may provide evidence (however, no proof) of microbial interactions.

For evaluation purposes we used PIDA and a SAGs-dataset. First, we wanted to evaluate the BBMO network via known interactions. Most interactions are unknown but a few have been recorded in the literature and gathered, for example, in the protist interaction database PIDA (manual curation) (Bjorbækmo *et al.*, 2019). The results of this attempt were poor. Only 29 (0.1%) association in the BBMO network have been found to correspond to eight interactions described in the literature of which 18 associations (corresponding to 2 interactions) were in a network-based environmental triplet (Chapter 5). Next, we employed Single Cell data but found little support for potential microbial interactions (results not shown in this thesis). Subsequently, we disregarded the analysis. However, as more single-cell data is and will be produced, it may be a valuable resource in the future. Thus, although literature-described interactions (PIDA) and SAGs were not enough to accurately evaluate our BBMO network, future extended versions and data will be promising evidence that a predicted edge is a true interaction.

Given that marine microbial interactions are barely known (Bjorbækmo *et al.*, 2019) and we lack comprehensive biological benchmark data, network construction evaluation is to a large extent carried out in silico (Faust, 2021). Similar to network construction evaluation lacking gold-standard datasets, we lacked a dataset to evaluate the different methods of detecting environmentally-driven edges (Chapter 5). Thus, to evaluate EnDED we needed a simulated dataset with known interactions. Such dataset required specific characteristics in order to evaluate all four methods (SP, OL, II, DPI) and their ensemble approach. The evaluation required temporal data with environmental perturbation generating indirect dependencies, i.e., changing environmental variables influencing the microorganisms. Public simulated data or approaches often do not consider the temporal dimension (e.g., (Berry & Widder, 2014; Röttjers & Faust, 2018)) or missed an environmental component (e.g., Berry & Widder, 2014) making them not applicable for our purposes. Thus, we simulated microbial abundances using an adjusted generalized Lotka-Volterra (gLV) model.

Generalized Lotka-Volterra models are popular to model microbial abundances despite several limitations (Gonze *et al.*, 2018). Moreover, environmental variables can be incorporated directly into the gLV (Dam *et al.*, 2016; Röttjers & Faust, 2018). Models in which all

microorganisms feed on the same nutrient are not suitable because, by definition, all microorganism would then have the ecological interaction of food competition. We adapted the gLV model and simulated datasets for our purposes. Our solution may be simple, but covered our most pressing needs: i) temporal dataset, ii) environmental factor influencing microorganisms but iii) microorganism not changing the environmental factor. The code is freely available and accessible for microbial ecologists.

For instance, our approach may be applicable for network construction evaluations since different approaches and data should be used (Faust, 2021). Faust (2021) argues that for most microbial communities it is not precisely known which processes shape microbial abundances. Further, evaluation should include a range of data simulation procedures incorporating various levels of noise because an evaluation relying on a single simulation method may favor tools with assumptions similar to the data simulation assumptions (Faust, 2021). For instance, several data generation methods have been used to evaluate eight microbial network inference tools (Weiss *et al.*, 2016). Tool development should be separated from evaluations, i.e., tool developers should be equipped with more heterogeneous benchmark data leading to tools performing in diverse settings (Faust, 2021), similar to the DREAM challenge for gene regulatory network inference (Marbach *et al.*, 2010). In accordance, heterogeneous benchmark data will benefit the systematic evaluation of methods to detect environmentally-driven indirect dependencies.

Conclusion

Currently, there is no best network construction method. Researchers should choose the tool that best handles specific characteristics of the dataset. Moreover, marine microbial studies lack gold-standard datasets (plural!) for proper benchmarking of network construction tools but also methods of indirect dependency detection. Simulating a dataset serving particular needs of a marine microbial system was outside of the thesis aims and remains subject to further research. However, our solution simulating temporal data with environmental influence served the purpose to evaluate the program EnDED. There may be more sophisticated tools to model microbial interactions but they are usually more complex and harder to interpret. Thus, our simple approach may serve microbial ecologists in specific investigations or developers in tool evaluations.

Each network elucidates a part of the whole

Lastly, there seem to be two main lines of research. One line of research aims to modulate microbial interactions through differential equations, which necessitates several assumptions, and (numerous) parameters to be defined beforehand. The other line of research uses microbial sequence abundances and determines associations between them, which are comprised through networks. Depending on the network construction method, there may be also some parameter setting needed.

Each dataset has its own characteristics, its own environment, its own story to tell. Each network can elucidate a part of the whole. The second line of research uses the microbial sequence abundances and aims to describe and discover what is there. The first line of research defines environment, growth rates, nutrient absorption rates, etc., and obtains a microbial system modeling the real world. We can learn from both.

The work of this thesis locates in the second line of research. I'm using the network-flashlight to allow a sneak-peak into the marine microbial ecosystem. If the entire marine microbial ecosystem is a puzzle, this PhD thesis will act as puzzle pieces on the quest to complete the picture. Our main aim disentangling environmental-effects in association networks acted as a broom cleaning up the dust to get a better view. Then, generating monthly subnetworks, i.e., adding a temporal dimension, we saw the network changing in time, like a movie. With the third project we left the comfort of our home (the BBMO data), and went to the wide world, we zoomed out of the picture. To see if patterns we found in one location (e.g., warm versus cold water), can be found in other locations. Similar to looking behind the scenes, we looked below the surface. We dived into a new world. The environment and microbial communities are different, and so are the networks.

We can imagine a meta-network which includes all possible microbial interactions between all microorganisms. Each ecosystem at each time point is a subnetwork of the meta-network. This subnetwork is not induced, i.e., the presence of microorganisms does not necessarily lead to the presence of an interaction between them. A similar idea was presented in previous work using the term metaweb to describe the regional pool of species and their potential interactions (Dunne, 2006; Poisot *et al.*, 2012), and a network drawn from this pool is called a realization (Poisot *et al.*, 2012). Thus, a realization is a subnetwork of the metaweb. Realizations from different times, locations or environmental conditions may be aggregated (or merged) to reconstruct the metaweb (Poisot *et al.*, 2012).

Conclusion

Here, we use microbial sequence abundances to determine associations and association networks to investigate real microbial ecosystems. As datasets and the samples' environments have their own characteristics, so do the networks. Subsequently, each network provides puzzle pieces of the bigger microbial world picture – it represents a potential subnetwork of the meta-network. Each network may provide a sneak-peak of and a step towards the meta-network.

Chapter 12 Further future perspectives

Classifying ecological interactions

Association networks are not equivalent to ecological networks. An association may or may not encode for a real interaction. 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). Further, an association encoding for a real interaction can represent different ecological interactions. Given two interaction partners, the outcome whether an interaction benefits, is disadvantageous, or does not have an effect (neutral) on one or both partners, results in six possibilities. Further classifications are possible, e.g., an antagonistic interaction could be due to predation or parasitism. However, the boundaries between the six categories can be blurred in real-life (Stat *et al.*, 2008) and the type of interaction may change between the same interaction partners, e.g., when transitioning from low to high-stress environments (Piccardi *et al.*, 2019). Thus, identifying ecological interactions is challenging.

Additional information may reduce the set of possible ecological interaction classifications. Given a suitable sampling periodicity in temporal data analysis, the network construction tool eLSA (Xia *et al.*, 2011, 2013) allows inferring time-delayed associations. Time-delay further characterizes an association and could potentially support ecological classification, e.g., a time-delayed association may point to predator-prey relationships. Furthermore, eLSA uses a dynamical programming approach allowing to infer associations over a subinterval of the time period. Imagine a scenario of an inferred negative association between two closely related microorganisms A and B. Microorganism A is an established microorganism appearing throughout the time-series but its sequence abundance reduces from the middle towards the end. Let microorganism B appear in the middle of the time series and from that point increase in its sequence abundance. Such a pattern may hint at a competition relationship. Further analysis investigating their resource requirement then may reveal that the two microorganisms share metabolic pathways for certain resources, which would strengthen a competition relationship hypothesis. However, in nature a vast number of microorganisms are present. Intra- and inter-species relationships take place. It may cause a third factor (environment or microorganisms) for a certain ecological interaction to take place. 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.

Classification of associations into ecological interaction will benefit from the integration of different datatypes allowing the construction of metabolic networks (Muller *et al.*, 2018).

Metabolic networks can infer functional redundancies and functional dependences to elucidate the ecological nature of associations. Functional dependence may point to symbiotic or parasitic relationships. Functional redundancy information may point to competition. Moreover, if the ecological nature of an association is known for one microorganism, it predicts ecological interactions of functional redundant organisms. A temporal analysis of marine microbial association networks constructive for three consecutive winters found that microbial taxa may change interaction partners over time (Lambert *et al.*, 2021). Given microorganism A interacts with two interchangeable partners B and C, which are functional redundant. If the ecological interaction between A and B is known, it may point to the same ecological interaction between B and C.

Some non-metabolic interactions, such as commensalism by niche engineering (e.g. a first organism creating an environment to allow a second organism to colonize) or predation cannot be predicted from inferred metabolic networks (Muller *et al.*, 2018). Thus, complimentary analyses remain valuable, such as association networks or using single-cell data. By definition, single cell data investigates single cells. Despite contamination, there may be biological reasons for finding genetic material of other organisms in a single cell. For instance, microorganisms with close physical contact may not get separated during cell isolation. Moreover, finding digestive material of one microorganism in the single-cell of another allows to infer predator-prey relationships. Thus, single cell sequencing provides further evidence on ecological interactions such as species-specific prey preferences and symbiotic interactions (Martinez-Garcia *et al.*, 2012). For instance, bacterial DNA appeared in 19% of 906 eukaryotic single amplified genomes (SAGs) from the Gulf of Main (GoM) and 48% of 792 SAGs from the Mediterranean Sea (MS) (Brown *et al.*, 2020). More viral DNA has been found in 51% GoM SAGs and 35% MS SAGs. The fraction of cells containing viral DNA varied among eukaryotic lineages and are 100% for *Picozoa* and *Choanozoa*, which also contained significantly higher numbers of viral sequences than other identified taxa (Brown *et al.*, 2020). Brown et al. (2020) conclude that predation on free viral particles contributed to the observed patterns.

Conclusion

This thesis aimed to disentangle microbial associations to provide valuable interaction hypotheses. However, interactions are diverse and identifying their ecological nature challenging. Several approaches are available but the best classifications may be reached when focusing on few associations to gather further evidence in biological-driven investigations. Providing life scientist with a selection of best interaction hypotheses will benefit their work. Thus, employing different data types into marine microbial studies may elucidate the potential ecological nature of microbial associations.

Additional graph-theoretic approaches

After disentangling association networks, microbial investigations can benefit from employing other graph theoretic approaches.

Microbial network-based roles

While we disregarded local network patterns by using global network metrics in Chapter 6, we included them in Chapter 7. We used the local topological metric based on graphlets (Pržulj *et al.*, 2004). Further analyses diving deeper into the study of specific microorganisms may benefit from extending the node degree to the graphlet degree vector, i.e., counting the number of orbits a node touches. For instance, quantifying a nodes' local connection patterns over time may allow inferring seasonal microorganisms via recurring connection patterns. Such a network-based approach would complement the detection of seasonal microorganisms based on sequence abundances (Giner *et al.*, 2019). Furthermore, characterizing nodes via their graphlet degree distribution allows comparing nodes in two or more networks. Two nodes with similar roles may be great candidates for microorganisms playing similar roles in an ecosystem, which may provide a hypothesis about potential functional redundant microorganisms. Changing roles may indicate system relevant changes. For instance, changing network-based roles was used to identify cancerous genes (Malod-Dognin *et al.*, 2019).

Moreover, using local patterns (e.g., graphlets) to investigate microbial network-based roles in monthly or sample-specific networks may identify fixed and adaptive microorganisms. Fixed microorganisms preserve their network-based roles (they appear in the same orbits), while adaptive microorganisms have different roles in different networks, i.e., the role changes (they appear in different orbits). Interactions partners may change in time for both, fixed and adaptive microorganisms, indicating the possibility to adapt to changing environment or different, potentially functional redundant, partners. Prevalent microorganisms appearing throughout the year may maintain relationships to other prevalent microorganisms or they may connect to summer as well as winter microorganisms. In the latter case, such microorganisms may be potential candidates involved or aiding in the transition between seasons, e.g., from summer to winter when the network reassembles such as we found with the temporal BBMO network. Relationships that are maintained are potentially essential associations, core associations.

Identifying similar roles is used for aligning networks (network alignment). Finding microorganisms with similar roles in the network may potentially translate into similar functional roles in the ecosystem.

Network Alignment

Network alignments aim to infer similar regions between networks. Specifically, network alignments aim to find a node mapping of topologically or functionally similar regions between

the networks (Vijayan *et al.*, 2017). Network alignments have a wide range of applications and there are several proposed approaches, see a comparison in (Trung *et al.*, 2020). Aligning static networks of different marine microbial association networks may be promising. Another possible application would be to transfer knowledge from (better) known systems, e.g., using a network alignment between a marine microbial system and the human gut microbiome or other (non-microbial) systems. Numerous network alignment tools and approaches exist. For example, homogeneous network alignment assumes that nodes and edges are of the same type. In contrast, heterogenous network alignment (Gu *et al.*, 2018) may be suitable if association networks include different players, e.g., aligning eukaryotes with eukaryotes and prokaryotes with prokaryotes. Instead of aligning single static networks, temporal networks can be aligned (Vijayan *et al.*, 2017; Vijayan & Milenković, 2018; Aparício *et al.*, 2019). Several network aligners have been compared and some unified into the tool Ulign (Malod-Dognin *et al.*, 2017). The previous study proposed a gain of additional biological insights by aligning all available data types collectively rather than any particular data type in isolation from others (Malod-Dognin *et al.*, 2017). Future investigations may shift from pairwise to multiple network alignments despite a recent comparison indicating that the former is often better than the latter depending on the choice of evaluation test (Vijayan *et al.*, 2020).

Beyond simple networks

Our networks were simple. A further adjustment to the BBMO temporal network may include associations between layers (time-delayed associations). Moreover, we used microbial associations based on 16S/18S rRNA data but the integration of different (e.g., omics) datatypes may provide complementary information. A plethora of strategies allow integrating multi-omics datasets (Bersanelli *et al.*, 2016; Huang *et al.*, 2017), e.g., combining networks of available data-types as layers of a multi-layer network (Bianconi, 2018) or fusing them into one network (Wang *et al.*, 2014; Malod-Dognin *et al.*, 2019). Finally, going beyond pairwise associations accounts for interactions between three or more interaction partners and necessitates a generalized graph, so-called hypergraphs. Hypergraphs allow edges between more than two nodes (Golubski *et al.*, 2016). Although identifying true interactions between two partners is already challenging, including non-pairwise interactions seems promising. A first step may be to account for a third interaction partner, e.g., a mediator (Ai *et al.*, 2019). However, the mediator approach necessitates triplets, which are three pairwise associations. Thus, including non-pairwise associations requires the adjustment of association detection but also data structures, algorithms, and visualizations to handle, store, and analyze such associations. For instance, while pairwise interactions can be displayed as adjacency matrix or incidence matrix, non-pairwise networks will require the incidence matrix, which may require much memory. Thus, sparse hypergraphs may be better stored via an adapted edge list providing for each edge (row) a list of incident nodes.

Subsequently, network-tools need to be adapted to handle non-pairwise edges. Considering visualization, an additional type of node may be added for each edge as in (Ai *et al.*, 2019), or colored planes are used, i.e., multi-node edges may be visualized via polygons (Figure 30). Large networks may be visualized through their incidence matrix as heat maps.

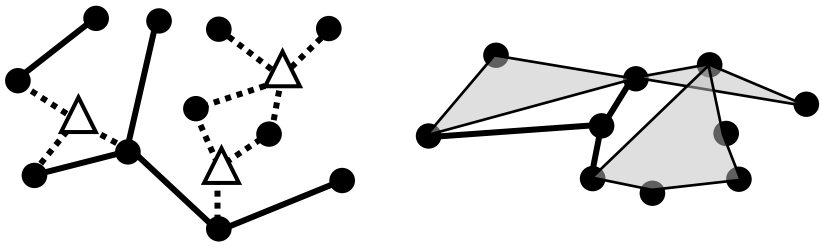


Figure 30: **Visualization hypergraphs**. Non-pairwise edges in hypergraphs may be represented via an additional node (triangle in the left hypergraph) or a polygon (gray polygons in the right network).

Conclusion

The graph-theoretic toolbox provides further valuable network-based approaches for marine microbial interaction studies. Determining the graphlet degree vector (network-based roles) may aid identifying functional redundancy and dependence. Aligning networks of different systems allows transferring knowledge from a well-studied system to a less known system. Different datatypes can be incorporated using a multilayer network or network fusion. Finally, going beyond pairwise interactions accounts for interactions of three or more partners (hypergraphs).

Chapter 13 All conclusions

First, I gather the conclusions stated previously, then I conclude the thesis.

Gathered conclusions

Chapter 5

⇒ We present 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.

Chapter 6

⇒ Incorporating the temporal dimension in the microbial association analysis unveiled multiple patterns that often remain hidden when using static networks. We developed a post-network-construction approach to generate a temporal network from a single static network that represents a step forward for disentangling the temporal nature of microbial associations. Yet, this approach has limitations, such as the monthly sampling frequency in our study. Using a higher sampling frequency would be the main solution. Investigating a coastal marine microbial ecosystem over ten years revealed a one-year-periodicity in the network topology. The temporal architecture was not stochastic, but displayed a modest amount of recurrence over time, especially in winter. Altogether, our approach allows comparing (sub)networks across spatiotemporal scales. Future efforts to understand the ocean microbiome should consider the dynamics of microbial interactions as these can be basis of ecosystem function.

Chapter 7

⇒ Our network-based exploration disentangles the spatial distribution of associations of the global ocean microbiome, from top to bottom layers, suggesting both global and regional interactions. Our analysis demonstrated the change of network topology across vertical (water column) and horizontal (different regions) dimensions of the ocean. Furthermore, our results indicate that associations have specific spatial distributions that are not just mirroring ASV distributions.

⇒ EnDED allows indirect dependency detection using network-based environmental triplets (generated during network construction) and artificially-generated triplets (generated after network construction). We quantified more environmentally-driven edges using artificially-generated triplets, especially for nutrients and the total chlorophyll-a concentration but only slightly more for temperature and day-length.

⇒ We showed how EnDED performed on the tools that we have previously used on Malaspina Surface data but other tools could have been included. Here, we determined how EnDED performs on associations detected via three different methods. Our results indicated that all tools were prone to indirect dependencies, at least to a minor fraction. We found less indirect dependencies among associations inferred by all three tools than associations inferred by one or two methods. Future systematic network construction tool benchmarking may include the quantification of environmentally-driven edges for which EnDED provides several methods.

⇒ EnDED integrates well in filter strategies, e.g., when determining core associations. Here, we extended the environmental factors by cell-count data. Cell-count data appears to be a valuable addition to detect environmentally-driven associations, i.e., indirect dependencies among microbial associations that are due to biotic factors. The number of heterotrophic prokaryotes, *Synechococcus*, and heterotrophic nanoflagellates cells infer much more negative than positive edges, similarly to other environmental factors. In contrast, the number of photosynthetic nanoflagellates cells remove a similar fraction of positive and negative associations. In sum, cell-count data are a valuable extension to the detection of environmentally-driven edges, especially positive associations.

⇒ The results indicate, that it is still worth considering general microbial triplets in future studies. While we investigated two types of microorganisms (phytoplankton affecting bacterial associations), future investigations should also consider triplets of the same type, e.g., three connected bacterial OTUs.

Chapter 9

⇒ A higher fraction of negative than positive associations was found to be environmentally-driven. Negative association may be true negative interactions or represent contrary environmental preference. Our results indicate that most negative associations in association networks may rather represent the latter, i.e., environmental preferences rather than true interactions.

⇒ Various factors may cause indirect dependencies in association networks such as abiotic factors (e.g., temperature), nutrients, but also biotic factors (e.g., specific

microorganisms, other living organisms or viruses). Thus, on one hand, future sampling should expand metadata collection in order to account for (more) abiotic and biotic factors that could explain and identify indirect dependencies. On the other hand, an updated version of EnDED should allow to investigate any triplet, e.g., microbial triplets.

⇒ The effect of environmentally-driven edges is prominent in association networks. All tested network construction tools were prone to effects of indirect dependencies but the extent of specific environmental drivers differed among tools. Thus, comparing environmental-drivers in microbial ecosystems will require same sampling, environmental data, filter and network construction strategies.

Chapter 10

⇒ Incorporating environmental data as additional nodes during network construction may miss important associations between environmental factors and microorganisms. Instead, it may be valuable to use different approaches to incorporate environmental data, e.g., inferring environmentally-driven associations via artificially-generated triplets and elucidating environmental influence on network architecture by relating environmental factors with network properties.

⇒ Sample-specific (including monthly) subnetworks provide a step forward towards disentangling the effects of environmental variables on network metrics and network architecture.

⇒ Focusing on the most significant and strongest associations may not be sufficient to select candidate interactions. Quantifying recurring associations may be promising and several approaches can be done: i) recurring associations using different network construction tools, ii) recurring associations using different datasets, iii) temporal recurrence, and iv) spatial recurrence. Quantifying microbial association recurrence may be a step towards identifying a core network, i.e., associations that are preserved across spatial and temporal scales.

⇒ Associations may be more repeatable at colder versus warmer waters, but it remains to be tested using a higher sampling frequency allowing the construction of season-specific networks over several years and depth-specific networks of one location, respectively. Moreover, such season-specific networks are the next step to test if recurring annual patterns in network architecture may be due to similar environmental variables. In accordance, depth-specific patterns and changes (up- and down-ward trends) may be better studied using networks constructed per depth and region instead of using subnetworks derived from a single static network. Yet, our approaches represent the first

steps in elucidating similar and different network-based patterns between seasons, depths, and between the ocean and the sea (here Mediterranean Sea).

Chapter 11

⇒ Future studies with higher sampling frequency may be able to construct networks within a month (monthly networks instead of subnetworks) and location (location-specific networks instead of subnetworks). Although our approaches are a good starting point that allow us to move forward, they have limitations suggesting caution when making biological interpretations from the temporal network and sample-specific subnetworks.

⇒ Currently, there is no best network construction method. Researchers should choose the tool that best handles specific characteristics of the dataset. Moreover, marine microbial studies lack gold-standard datasets (plural!) for proper benchmarking of network construction tools but also methods of indirect dependency detection. Simulating a dataset serving particular needs of a marine microbial system was outside of the thesis aims and remains subject to further research. However, our solution simulating temporal data with environmental influence served the purpose to evaluate the program EnDED. There may be more sophisticated tools to model microbial interactions but they are usually more complex and harder to interpret. Thus, our simple approach may serve microbial ecologists in specific investigations or developers in tool evaluations.

⇒ Here, we use microbial sequence abundances to determine associations and association networks to investigate real microbial ecosystems. As datasets and the samples' environments have their own characteristics, so do the networks. Subsequently, each network provides puzzle pieces of the bigger microbial world picture – it represents a potential subnetwork of the meta-network. Each network may provide a sneak-peak of and a step towards the meta-network.

Chapter 12

⇒ This thesis aimed to disentangle microbial associations to provide valuable interaction hypotheses. However, interactions are diverse and identifying their ecological nature challenging. Several approaches are available but the best classifications may be reached when focusing on few associations to gather further evidence in biological-driven investigations. Providing life scientist with a selection of best interaction hypotheses will benefit their work. Thus, employing different data types into marine microbial studies may elucidate the potential ecological nature of microbial associations.

⇒ The graph-theoretic toolbox provides further valuable network-based approaches for marine microbial interaction studies. Determining the graphlet degree vector (network-based roles) may aid identifying functional redundancy and dependence. Aligning

networks of different systems allows transferring knowledge from a well-studied system to a less known system. Different datatypes can be incorporated using a multilayer network or network fusion. Finally, going beyond pairwise interactions accounts for interactions of three or more partners (hypergraphs).

Thesis conclusion

Investigating microbial interactions is challenging due to microbial sampling problems and the challenges before, during, and after inferring microbial interactions through association networks. Although they do not provide complete information on the interactions inside the system, the network provides a system view, which has value on itself. Thus, networks are a valuable tool given their advantages. Subsequently, improving inferred networks before down-stream analysis will benefit microbial interaction studies. Here, the thesis presented our main ideas using network-based approaches to help disentangle microbial interactions. They resulted in the following conclusions.

- ⇒ EnDED aims to disentangle true ecological interactions from environmentally-driven associations. It can be applied with abiotic environmental factors, nutrients, but also cell-counts or specific microorganisms. Further, it may be promising to omit environmental factors during network construction but employ them post network construction.
- ⇒ We learnt from simulated networks, that a cut-off level on the association score is not sufficient to separate true from false interactions.
- ⇒ Other filter strategies should be applied and EnDED represents one puzzle piece, which may be followed by quantifying association recurrence. For that, current datasets may not provide sufficient numbers of samples but our approach generating monthly or sample-specific subnetworks from a single static network is a step forward allowing new perspectives on the dynamics of microbial ecosystems.
- ⇒ The main environmental drivers of indirect dependencies in the BBMO network (temporal data including 120 months of ten years) were Temperature and Daylength, while nutrients were the main drivers in the global network (spatial data including 397 samples covering the global ocean and Mediterranean Sea from the surface to the deep ocean).
- ⇒ The fraction of environmentally-driven associations among negative microbial associations increased rapidly with the number of network-based environmental triplets. In accordance, when adjusting the approach using artificially-generated triplets, we found a higher fraction of negative environmentally-driven edges. Similarly, a higher fraction of negative than positive edges was detected to be environmentally-driven when considering cell-counts or phytoplankton taxa.

⇒ Associations may represent permanent, temporary, or seasonal associations. Networks constructed for each specific period (hour, day, week, or month) allow their detection but such sampling over longer periods of time is costly. Our approach allows determining monthly subnetworks, which provided a step forward into investigating dynamics at the model marine microbial ecosystem at the BBMO from a network perspective allowing the generation of hypothesis of temporal interaction patterns.

⇒ Associations could be global or regional. Similar to temporal data, spatial data usually provides one or (too) few samples per location to construct location-specific networks. Our approach allows determining sample-specific subnetworks, which is a step towards investigating the biogeography of marine microbial interactions allowing the generation of hypotheses of spatial, e.g., depth-related, patterns.

⇒ It is known that most deep ocean ASVs already appeared in upper layers. We expected to find similar results for associations. In contrast, most associations in the mesopelagic (81.77-90.90%) and bathypelagic (43.54-72.71%) appeared for the first time in these layers in the five ocean basins. In the MS, it was 71.24% and 22.44%, respectively.

⇒ This thesis provided several methodologies to aid marine microbial network-based investigations.

References

ACINAS, S.G., SÁNCHEZ, P., SALAZAR, G., CORNEJO-CASTILLO, F.M., SEBASTIÁN, M., LOGARES, R., ROYO-LLONCH, M., PAOLI, L., SUNAGAWA, S., HINGAMP, P., OGATA, H., LIMA-MENDEZ, G., ROUX, S., GONZÁLEZ, J.M., ARRIETA, J.M., ALAM, I.S., KAMAU, A., BOWLER, C., RAES, J., PESANT, S., BORK, P., AGUSTÍ, S., GOJOBORI, T., VAQUÉ, D., SULLIVAN, M.B., PEDRÓS-ALIÓ, C., MASSANA, R., DUARTE, C.M., & GASOL, J.M. (2021) Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Communications Biology*, **4**, 604.

AGUSTI, S., GONZÁLEZ-GORDILLO, J.I., VAQUÉ, D., ESTRADA, M., CEREZO, M.I., SALAZAR, G., GASOL, J.M., & DUARTE, C.M. (2015) Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. *Nature Communications*, **6**, 7608.

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*.

AITCHISON, J. (1986) *The statistical analysis of compositional data*. London; New York: Chapman and Hall.

ALBANESE, D., RICCADONNA, S., DONATI, C., & FRANCESCHI, P. (2018) A practical tool for maximal information coefficient analysis. *GigaScience*, **7**.

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

ALM, E. & ARKIN, A.P. (2003) Biological networks. *Current Opinion in Structural Biology*, **13**, 193–202.

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.

AMARAL, L.A.N. & OTTINO, J.M. (2004) Complex networks. *The European Physical Journal B*, **38**, 147–162.

APARÍCIO, D., RIBEIRO, P., MILENKOVIĆ, T., & SILVA, F. (2019) Temporal network alignment via GoT-WAVE. *Bioinformatics*, **35**, 3527–3529.

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.

ARANDIA-GOROSTIDI, N., KRABBERØD, A.K., LOGARES, R., DEUTSCHMANN, I.M., SCHAREK, R., MORÁN, X.A.G., GONZÁLEZ, F., ALONSO-SÁEZ, L. (in preparation) Biotic interactions between phytoplankton and heterotrophic bacteria dominate microbial seasonal dynamic in coastal ocean waters.

ARÍSTEGUI, J., GASOL, J.M., DUARTE, C.M., & HERNDLD, G.J. (2009) Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, **54**, 1501–1529.

AULADELL, A., BARBERÁN, A., LOGARES, R., GARCÉS, E., GASOL, J.M., & FERRERA, I. (2020) Seasonal niche differentiation between evolutionary closely related marine bacteria. *bioRxiv*, 2020.12.17.423265.

BAKER, G.C., SMITH, J.J., & COWAN, D.A. (2003) Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods*, **55**, 541–555.

BALDAUF, S.L. (2008) An overview of the phylogeny and diversity of eukaryotes. *Journal of Systematics and Evolution*, **46**, 263.

BANERJEE, S., SCHLAEPPI, K., & VAN DER HEIJDEN, M.G.A. (2018) Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology*, **16**, 567–576.

BARABÁSI, A.-L. (2003) *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. New York: Plume Editors.

BARABÁSI, A.-L. & ALBERT, R. (1999) Emergence of Scaling in Random Networks. *Science*, **286**, 509.

BARBERÁN, A., BATES, S.T., CASAMAYOR, E.O., & FIERER, N. (2012) Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal*, **6**, 343–351.

BAR-ON, Y.M. & MILO, R. (2019) The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet. *Cell*, **179**, 1451–1454.

BAR-ON, Y.M., PHILLIPS, R., & MILO, R. (2018) The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, **115**, 6506–6511.

BARRACLOUGH, T.G. (2015) How Do Species Interactions Affect Evolutionary Dynamics Across Whole Communities? *Annu. Rev. Ecol. Evol. Syst.*, **46**, 25–48.

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.

BASTIAN, M., HEYMANN, S., & JACOMY, M. (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. *ICWSM*, **3**.

BEKKER, A., HOLLAND, H.D., WANG, P.-L., RUMBLE, D., STEIN, H.J., HANNAH, J.L., COETZEE, L.L., & BEUKES, N.J. (2004) Dating the rise of atmospheric oxygen. *Nature*, **427**, 117–120.

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.

BERSANELLI, M., MOSCA, E., REMONDINI, D., GIAMPIERI, E., SALA, C., CASTELLANI, G., & MILANESI, L. (2016) Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*, **17**, S15.

BHARTI, R. & GRIMM, D.G. (2021) Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, **22**, 178–193.

BIANCONI, G. (2018) *Multilayer networks: structure and function*. Oxford university press.

BIGGS, N., LLOYD, E.K., & WILSON, R.J. (1986) *Graph Theory, 1736-1936*. Oxford University Press.

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.

BLONDEL, V.D., GUILLAUME, J.-L., LAMBIOTTE, R., & LEFEBVRE, E. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**, P10008.

BLONDER, B., WEY, T.W., DORNHAUS, A., JAMES, R., & SIH, A. (2012) Temporal dynamics and network analysis. *Methods in Ecology and Evolution*, **3**, 958–972.

BOEUF, D., EDWARDS, B.R., EPPLEY, J.M., HU, S.K., POFF, K.E., ROMANO, A.E., CARON, D.A., KARL, D.M., & DELONG, E.F. (2019) Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proc Natl Acad Sci USA*, **116**, 11824.

BORENSTEIN, E. & FELDMAN, M.W. (2009) Topological Signatures of Species Interactions in Metabolic Networks. *Journal of Computational Biology*, **16**, 191–200.

BOYER, T.P., ANTONOV, J.I., BARANOVA, O.K., GARCIA, H.E., JOHNSON, D.R., MISHONOV, A.V., O'BRIEN, T.D., SEIDOV, D., 1948-, SMOLYAR, I. (Igor), ZWENG, M.M., PAVER, C.R., LOCARNINI, R.A., REAGAN, J.R., FORGY, C. (Carla), GRODSKY, A., & LEVITUS, S. (2013) World ocean database 2013. NOAA atlas NESDIS ; 72, DOI: 10.7289/V5NZ85MT.

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.

BROWN, J.M., LABONTÉ, J.M., BROWN, J., RECORD, N.R., POULTON, N.J., SIERACKI, M.E., LOGARES, R., & STEPANAUSKAS, R. (2020) Single Cell Genomics Reveals Viruses Consumed by Marine Protists. *Frontiers in Microbiology*, **11**, 2317.

BUNSE, C. & PINHASSI, J. (2017) Marine Bacterioplankton Seasonal Succession Dynamics. *Trends in Microbiology*, **25**, 494–505.

CALLAHAN, B.J., MCMURDIE, P.J., & HOLMES, S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, **11**, 2639–2643.

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.

CHAFFRON, S., DELAGE, E., BUDINICH, M., VINTACHE, D., HENRY, N., NEF, C., ARDYNA, M., ZAYED, A.A., JUNGER, P.C., GALAND, P.E., LOVEJOY, C., MURRAY, A., SARMENTO, H., ACINAS, S., BABIN, M., IUDICONE, D., JAILLON, O., KARSENTI, E., WINCKER, P., KARP-BOSS, L., SULLIVAN, M.B., BOWLER, C., DE VARGAS, C., & EVEILLARD, D. (2020) Environmental vulnerability of the global ocean plankton community interactome. *bioRxiv*, 2020.11.09.375295.

CHAFFRON, S., REHRAUER, H., PERNTHALER, J., & VON MERING, C. (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome research*, **20**, 947–959.

CHAMBOUVET, A., MORIN, P., MARIE, D., & GUILLOU, L. (2008) Control of Toxic Marine Dinoflagellate Blooms by Serial Parasitic Killers. *Science*, **322**, 1254.

CHOW, C.-E.T., KIM, D.Y., SACHDEVA, R., CARON, D.A., & FUHRMAN, J.A. (2014) Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. *The ISME Journal*, **8**, 816–829.

CHOW, C.-E.T., SACHDEVA, R., CRAM, J.A., STEELE, J.A., NEEDHAM, D.M., PATEL, A., PARADA, A.E., & FUHRMAN, J.A. (2013) Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. *The ISME Journal*, **7**, 2259–2273.

CONNOR, N., BARBERÁN, A., & CLAUSET, A. (2017) Using null models to infer microbial co-occurrence networks. *PLOS ONE*, **12**, 1–23.

COUGOUL, A., BAILLY, X., VOURC'H, G., & GASQUI, P. (2019) Rarity of microbial species: In search of reliable associations. *PLOS ONE*, **14**, e0200458.

COUTINHO, F.H., MEIRELLES, P.M., MOREIRA, A.P.B., PARANHOS, R.P., DUTILH, B.E., & THOMPSON, F.L. (2015) Niche distribution and influence of environmental parameters in marine microbial communities: a systematic review. *PeerJ*, **3**, e1008.

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

CRAM, J.A., CHOW, C.-E.T., SACHDEVA, R., NEEDHAM, D.M., PARADA, A.E., STEELE, J.A., & FUHRMAN, J.A. (2015) Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *The ISME Journal*, **9**, 563–580.

CRAM, J.A., XIA, L.C., NEEDHAM, D.M., SACHDEVA, R., SUN, F., & FUHRMAN, J.A. (2015) Cross-depth analysis of marine bacterial networks suggests downward propagation of temporal changes. *The ISME Journal*, **9**, 2573–2586.

CSARDI, G. & NEPUSZ, T. (2006) The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

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.

DE VARGAS, C., AUDIC, S., HENRY, N., DECELLE, J., MAHÉ, F., LOGARES, R., LARA, E., BERNEY, C., LE BESCOT, N., PROBERT, I., CARMICHAEL, M., POULAIN, J., ROMAC, S., COLIN, S., AURY, J.-M., BITTNER, L., CHAFFRON, S., DUNTHORN, M., ENGELEN, S., FLEGONTOVA, O., GUIDI, L., HORÁK, A., JAILLON, O., LIMA-MENDEZ, G., LUKEŠ, J., MALVIYA, S., MORARD, R., MULOT, M., SCALCO, E., SIANO, R., VINCENT, F., ZINGONE,

A., DIMIER, C., PICHERAL, M., SEARSON, S., KANDELS-LEWIS, S., ACINAS, S.G., BORK, P., BOWLER, C., GORSKY, G., GRIMSLEY, N., HINGAMP, P., IUDICONE, D., NOT, F., OGATA, H., PESANT, S., RAES, J., SIERACKI, M.E., SPEICH, S., STEMMANN, L., SUNAGAWA, S., WEISSENBACH, J., WINCKER, P., & KARSENTI, E. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.

DEL GIORGIO, P.A. & DUARTE, C.M. (2002) Respiration in the open ocean. *Nature*, **420**, 379–384.

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

DEMAIN, A.L. & ADRIO, J.L. (2008) Contributions of microorganisms to industrial biology. *Molecular Biotechnology*, **38**, 41.

DEUTSCHMANN, I., KRABBERØD, A.K., BENITES, L.F., LATORRE, F., DELAGE, E., MARRASÉ, C., BALAGUÉ, V., GASOL, J.M., MASSANA, R., EVEILLARD, D., CHAFFRON, S., & LOGARES, R. (2021) Disentangling temporal associations in marine microbial networks. *Research Square*, DOI: 10.21203/rs.3.rs-404332/v1.

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

DEUTSCHMANN, I.M., LIMA-MENDEZ, G., KRABBERØD, A.K., RAES, J., VALLINA, S.M., FAUST, K., & LOGARES, R. (2020) Disentangling environmental effects in microbial association networks. *PREPRINT (Version1) available at Research Square*, DOI: 10.21203/rs.3.rs-57387/v1.

DIDIER, G., BRUN, C., & BAUDOT, A. (2015) Identifying communities from multiplex biological networks. *PeerJ*, **3**, e1525.

DIDIER, G., VALDEOLIVAS, A., & BAUDOT, A. (2018) Identifying communities from multiplex biological networks by randomized optimization of modularity [version 2; peer review: 3 approved, 1 approved with reservations]. *F1000Research*, **7**.

DUARTE, C.M. (2015) Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin*, **24**, 11–14.

DUNNE, J.A. (2006) The network structure of food webs. *Ecological networks: linking structure to dynamics in food webs*, 27–86.

DUPONT, C.L., CHAPPELL, D., LOGARES, R., & VILA-COSTA, M. (2010) A hitchhiker's guide to the new molecular toolbox for ecologists. *Eco-DAS VIII Symposium Proceedings*. pp. 17–29.

EDGAR, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, **10**, 996–998.

ENDO, H., BLANC-MATHIEU, R., LI, Y., SALAZAR, G., HENRY, N., LABADIE, K., DE VARGAS, C., SULLIVAN, M.B., BOWLER, C., WINCKER, P., KARP-BOSS, L., SUNAGAWA, S., & OGATA, H. (2020) Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature Ecology & Evolution*, **4**, 1639–1649.

ERDÖS, P. & RÉNYI, A. (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, **5**, 17–60.

ERLICH, H.A. (1989) *PCR technology*, vol. 246. Springer.

ESPEJO, R., MESTRE, G., POSTIGO, F., LUMBRERAS, S., RAMOS, A., HUANG, T., & BOMPARD, E. (2020) Exploiting graphlet decomposition to explain the structure of complex networks: the GHuST framework. *Scientific Reports*, **10**, 12884.

ESTRADA, E. & KNIGHT, P.A. (2015) *A first course in network theory*. Oxford University Press, USA.

ESTRADA, M. (1996) Primary production in the northwestern Mediterranean.

EULER, L. (1741) Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 128–140.

FALKOWSKI, P.G. (2015) *Life's engines: how microbes made Earth habitable*, vol. 24. Princeton University Press.

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. (2021) Open challenges for microbial network construction and analysis. *The ISME Journal*, DOI: 10.1038/s41396-021-01027-4.

FAUST, K., LAHTI, L., GONZE, D., VOS, W.M. de, & RAES, J. (2015) Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology*, **25**, 56–66.

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.

FERRERA, I., REÑÉ, A., FUNOSAS, D., CAMP, J., MASSANA, R., GASOL, J.M., & GARCÉS, E. (2020) Assessment of microbial plankton diversity as an ecological indicator in the NW Mediterranean coast. *Marine Pollution Bulletin*, **160**, 111691.

FIELD, C.B., BEHRENFELD, M.J., RANDERSON, J.T., & FALKOWSKI, P. (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, **281**, 237.

FLEISCHMANN, R., ADAMS, M., WHITE, O., CLAYTON, R., KIRKNESS, E., KERLAVAGE, A., BULT, C., TOMB, J., DOUGHERTY, B., MERRICK, J., & AL., *et al*. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496.

FREDRICKSON, A. & STEPHANOPOULOS, G. (1981) Microbial competition. *Science*, **213**, 972.

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

FRUCHTERMAN, T.M.J. & REINGOLD, E.M. (1991) Graph drawing by force-directed placement. *Software: Practice and Experience*, **21**, 1129–1164.

FUHRMAN, J.A., CRAM, J.A., & NEEDHAM, D.M. (2015) Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, **13**, 133–146.

GARCZAREK, L., GUYET, U., DORÉ, H., FARRANT, G.K., HOEBEKE, M., BRILLET-GUÉGUEN, L., BISCH, A., FERRIEUX, M., SILTANEN, J., CORRE, E., LE CORGUILLÉ, G., RATIN, M., PITT, F.D., OSTROWSKI, M., CONAN, M., SIEGEL, A., LABADIE, K., AURY, J.-M., WINCKER, P., SCANLAN, D.J., & PARTENSKY, F. (2021) Cyanorak v2.1: a scalable information system dedicated to the visualization and expert curation of marine and brackish picocyanobacteria genomes. *Nucleic Acids Research*, **49**, D667–D676.

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*.

GAST, R.J. & CARON, D.A. (1996) Molecular phylogeny of symbiotic dinoflagellates from planktonic foraminifera and radiolaria. *Molecular Biology and Evolution*, **13**, 1192–1197.

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.

GINER, C.R., PERNICE, M.C., BALAGUÉ, V., DUARTE, C.M., GASOL, J.M., LOGARES, R., & MASSANA, R. (2020) Marked changes in diversity and relative activity of picoeukaryotes with depth in the world ocean. *The ISME Journal*, **14**, 437–449.

GIRVAN, M. & NEWMAN, M.E.J. (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA*, **99**, 7821.

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.

GOLUBSKI, A.J., WESTLUND, E.E., VANDERMEER, J., & PASCUAL, M. (2016) Ecological Networks over the Edge: Hypergraph Trait-Mediated Indirect Interaction (TMII) Structure. *Trends in Ecology & Evolution*, **31**, 344–354.

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.

GU, S., JOHNSON, J., FAISAL, F.E., & MILENKOVIĆ, T. (2018) From homogeneous to heterogeneous network alignment via colored graphlets. *Scientific Reports*, **8**, 12524.

GU, Z., GU, L., EILS, R., SCHLESNER, M., & BRORS, B. (2014) circlize implements and enhances circular visualization in R. *Bioinformatics*, **30**, 2811–2812.

GUERRERO, R., PEDRÓS-ALIÓ, C., ESTEVE, I., MAS, J., CHASE, D., & MARGULIS, L. (1986) Predatory prokaryotes: Predation and primary consumption evolved in bacteria. *Proc Natl Acad Sci USA*, **83**, 2138.

GUIDI, L., CHAFFRON, S., BITTNER, L., EVEILLARD, D., LARHLIMI, A., ROUX, S., DARZI, Y., AUDIC, S., BERLINE, L., BRUM, J.R., COELHO, L.P., ESPINOZA, J.C.I., MALVIYA, S., SUNAGAWA, S., DIMIER, C., KANDELS-LEWIS, S., PICHERAL, M., POULAIN, J., SEARSON, S., STEMMANN, L., NOT, F., HINGAMP, P., SPEICH, S., FOLLOWS, M., KARP-BOSS, L., BOSS, E., OGATA, H., PESANT, S., WEISSENBACH, J., WINCKER, P., ACINAS, S.G., BORK, P., DE VARGAS, C., IUDICONE, D., SULLIVAN, M.B., RAES, J., KARSENTI, E., BOWLER, C., GORSKY, G., & TARA OCEANS CONSORTIUM COORDINATORS (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, **532**, 465–470.

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.

HALL, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.*, **210**, 1518.

HANSELL, D.A. & CARLSON, C.A. (1998) Deep-ocean gradients in the concentration of dissolved organic carbon. *Nature*, **395**, 263–266.

HASIN, Y., SELDIN, M., & LUSIS, A. (2017) Multi-omics approaches to disease. *Genome Biology*, **18**, 83.

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.

HERNANDEZ, D.J., DAVID, A.S., MENGES, E.S., SEARCY, C.A., & AFKHAMI, M.E. (2021) Environmental stress destabilizes microbial networks. *The ISME Journal*, DOI: 10.1038/s41396-020-00882-x.

HIRANO, H. & TAKEMOTO, K. (2019) Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics*, **20**, 329.

HOLM, S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

HUANG, S., CHAUDHARY, K., & GARMIRE, L.X. (2017) More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*, **8**, 84.

IBARBALZ, F.M., HENRY, N., BRANDÃO, M.C., MARTINI, S., BUSSENI, G., BYRNE, H., COELHO, L.P., ENDO, H., GASOL, J.M., GREGORY, A.C., MAHÉ, F., RIGONATO, J., ROYO-LLONCH, M., SALAZAR, G., SANZ-SÁEZ, I., SCALCO, E., SOVIADAN, D., ZAYED, A.A., ZINGONE, A., LABADIE, K., FERLAND, J., MAREC, C., KANDELS, S., PICHERAL, M., DIMIER, C., POULAIN, J., PISAREV, S., CARMICHAEL, M., PESANT, S., ACINAS, S.G., BABIN, M., BORK, P., BOSS, E., BOWLER, C., COCHRANE, G., VARGAS, C. de, FOLLOWS, M., GORSKY, G., GRIMSLEY, N., GUIDI, L., HINGAMP, P., IUDICONE, D., JAILLON, O., KANDELS, S., KARP-BOSS, L., KARSENTI, E., NOT, F., OGATA, H., PESANT, S., POULTON, N., RAES, J., SARDET, C., SPEICH, S., STEMMANN, L., SULLIVAN, M.B., SUNAGAWA, S., WINCKER, P., BABIN, M., BOSS, E., IUDICONE, D., JAILLON, O., ACINAS, S.G., OGATA, H., PELLETIER, E., STEMMANN, L., SULLIVAN, M.B., SUNAGAWA, S., BOPP, L., VARGAS, C. de, KARP-BOSS, L., WINCKER, P., LOMBARD, F., BOWLER, C., & ZINGER, L. (2019) Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, **179**, 1084-1097.e21.

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

JAMAKOVIC, A. & UHLIG, S. (2008) On the relationships between topological measures in real-world networks. *Networks & Heterogeneous Media*, **3**, 345–359.

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.

JARDILLIER, L., ZUBKOV, M.V., PEARMAN, J., & SCANLAN, D.J. (2010) Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *The ISME Journal*, **4**, 1180–1192.

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.

KARSENTI, E., ACINAS, S.G., BORK, P., BOWLER, C., DE VARGAS, C., RAES, J., SULLIVAN, M., ARENDT, D., BENZONI, F., CLAVERIE, J.-M., FOLLOWS, M., GORSKY, G., HINGAMP, P., IUDICONE, D., JAILLON, O., KANDELS-LEWIS, S., KRZIC, U., NOT, F., OGATA, H., PESANT, S., REYNAUD, E.G., SARDET, C., SIERACKI, M.E., SPEICH, S., VELAYOUDON, D., WEISSENBACH, J., WINCKER, P., & THE TARA OCEANS CONSORTIUM (2011) A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biology*, **9**, e1001177.

KEELING, P.J. & CAMPO, J. del (2017) Marine Protists Are Not Just Big Bacteria. *Current Biology*, **27**, R541–R549.

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.

KHANIN, R. & WIT, E. (2006) How Scale-Free Are Biological Networks. *Journal of Computational Biology*, **13**, 810–818.

KITANO, H. (2004) Biological robustness. *Nature Reviews Genetics*, **5**, 826–837.

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

KOUTROULI, M., KARATZAS, E., PAEZ-ESPINO, D., & PAVLOPOULOS, G.A. (2020) A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology*, **8**, 34.

KOWALLIK, K.V. & MARTIN, W.F. (2021) The origin of symbiogenesis: An annotated English translation of Mereschkowsky's 1910 paper on the theory of two plasma lineages. *Biosystems*, **199**, 104281.

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.

KRABBERØD, A.K., DEUTSCHMANN, I.M., BJORBÆKMO, M.F.M., BALAGUÉ, V., GINER, C.R., FERRERA, I., GARCÉS, E., MASSANA, R., GASOL, J.M., & LOGARES, R. (2021) Long-term patterns of an interconnected core marine microbiota. *bioRxiv*, 2021.03.18.435965.

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*.

LAMBERT, S., LOZANO, J.-C., BOUGET, F.-Y., & GALAND, P.E. (2021) Seasonal marine microorganisms change neighbours under contrasting environmental conditions. *Environmental Microbiology*, **23**, 2592–2604.

LAMBERT, S., TRAGIN, M., LOZANO, J.-C., GHIGLIONE, J.-F., VAULOT, D., BOUGET, F.-Y., & GALAND, P.E. (2019) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *The ISME Journal*, **13**, 388.

LANGFELDER, P. & HORVATH, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

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.

LATOUCHE, P., BIRMELÉ, E., & AMBROISE, C. (2011) Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, **5**, 309–336.

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.

LEWIS, W.H., TAHON, G., GEESINK, P., SOUSA, D.Z., & ETTEMA, T.J.G. (2020) Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*, DOI: 10.1038/s41579-020-00458-8.

LI, K.-C. (2002) Genome-wide coexpression dynamics: Theory and application. *Proc Natl Acad Sci USA*, **99**, 16875–16880.

LI, C., LIM, K.M.K., CHNG, K.R., & NAGARAJAN, N. (2016) Predicting microbial interactions through computational approaches. *Methods*.

LI, W.K.W. (1994) Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. *Limnology and Oceanography*, **39**, 169–175.

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.

LINDSTRÖM, E.S. & LANGENHEDER, S. (2012) Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, **4**, 1–9.

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

LOGARES, R., DEUTSCHMANN, I.M., JUNGER, P.C., GINER, C.R., KRABBERØD, A.K., SCHMIDT, T.S.B., RUBINAT-RIPOLL, L., MESTRE, M., SALAZAR, G., RUIZ-GONZÁLEZ, C., SEBASTIÁN, M., DE VARGAS, C., ACINAS, S.G., DUARTE, C.M., GASOL, J.M., &

MASSANA, R. (2020) Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome*, **8**, 55.
LOGARES, R., HAVERKAMP, T.H.A., KUMAR, S., LANZÉN, A., NEDERBRAGT, A.J., QUINCE, C., & KAUSERUD, H. (2012) Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches. *Journal of Microbiological Methods*, **91**, 106–113.
LÓPEZ-GARCÍA, P., EME, L., & MOREIRA, D. (2017) Symbiosis in eukaryotic evolution. *Journal of Theoretical Biology*, **434**, 20–33.
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.
MACAULAY, I.C. & VOET, T. (2014) Single Cell Genomics: Advances and Future Perspectives. *PLOS Genetics*, **10**, e1004126.
MALOD-DOGNIN, N., BAN, K., & PRŽULJ, N. (2017) Unified Alignment of Protein-Protein Interaction Networks. *Scientific Reports*, **7**, 953.
MALOD-DOGNIN, N., PETSCHNIGG, J., WINDELS, S.F.L., POVH, J., HEMINGWAY, H., KETTELER, R., & PRŽULJ, N. (2019) Towards a data-integrated cell. *Nature Communications*, **10**, 805.
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.
MARBACH, D., PRILL, R.J., SCHAFFTER, T., MATTIUSSI, C., FLOREANO, D., & STOLOVITZKY, G. (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, **107**, 6286–6291.
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.
MARTIN, B.D. & SCHWAB, E. (2013) Current usage of symbiosis and associated terminology. *International Journal of Biology*, **5**, 32.
MARTIN, W. & KOWALLIK, K. (1999) Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche.' *European Journal of Phycology*, **34**, 287–295.
MARTINEZ-GARCIA, M., BRAZEL, D., POULTON, N.J., SWAN, B.K., GOMEZ, M.L., MASLAND, D., SIERACKI, M.E., & STEPANAUSKAS, R. (2012) Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *The ISME Journal*, **6**, 703–707.
MARTÍNEZ-PÉREZ, A.M., OSTERHOLZ, H., NIETO-CID, M., ÁLVAREZ, M., DITTMAR, T., & ÁLVAREZ-SALGADO, X.A. (2017) Molecular composition of dissolved organic matter in the Mediterranean Sea. *Limnology and Oceanography*, **62**, 2699–2712.
MARTÍN-FERNÁNDEZ, J.A., BARCELÓ-VIDAL, C., & PAWLOWSKY-GLAHN, V. (2003) Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology*, **35**, 253–278.
MARTÍN-FERNÁNDEZ, J.A., HRON, K., TEMPL, M., FILZMOSER, P., & PALAREA-ALBALADEJO, J. (2012) Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics & Data Analysis*, **56**, 2688–2704.
MASSANA, R. (2011) Eukaryotic Picoplankton in Surface Oceans. *Annu. Rev. Microbiol.*, **65**, 91–110.
MASSANA, R. & LOGARES, R. (2013) Eukaryotic versus prokaryotic marine picoplankton ecology. *Environmental Microbiology*, **15**, 1254–1261.
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.

McInnes, L., Healy, J., & Astels, S. (2017) hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, **2**, 205.

McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018) UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, **3**, 861.

McMurdie, P.J. & Holmes, S. (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*, **10**, e1003531.

McNichol, J., Berube, P.M., Biller, S.J., & Fuhrman, J.A. (2020) Evaluating and Improving SSU rRNA PCR Primer Coverage via Metagenomes from Global Ocean Surveys. *bioRxiv*, 2020.11.09.375543.

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.

Mereschkowsky, C. (1905) Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biologisches Centralblatt*, **25**, 293–604.

Mereschkowsky, C. (1910) Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. *Biologisches Centralblatt*, **30**, 278–288.

Mestre, M., Höfer, J., Sala, M.M., & Gasol, J.M. (2020) Seasonal Variation of Bacterial Diversity Along the Marine Particulate Matter Continuum. *Frontiers in Microbiology*, **11**, 1590.

Mestre, M., Ruiz-González, C., Logares, R., Duarte, C.M., Gasol, J.M., & Sala, M.M. (2018) Sinking particles promote vertical connectivity in the ocean microbiome. *Proc Natl Acad Sci USA*, **115**, E6799.

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.

Milici, M., Deng, Z.-L., Tomasch, J., Decelle, J., Wos-Oxley, M.L., Wang, H., Jáuregui, R., Plumeier, I., Giebel, H.-A., Badewien, T.H., Wurst, M., Pieper, D.H., Simon, M., & Wagner-Döbler, I. (2016) Co-occurrence Analysis of Microbial Taxa in the Atlantic Ocean Reveals High Connectivity in the Free-Living Bacterioplankton. *Frontiers in Microbiology*, **7**, 649.

Mori, A.S., Isbell, F., & Seidl, R. (2018) β-Diversity, Community Assembly, and Ecosystem Functioning. *Trends in Ecology & Evolution*, **33**, 549–564.

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

Muller, E.E.L., Faust, K., Widder, S., Herold, M., Martínez Arbas, S., & Wilmes, P. (2018) Using metabolic networks to resolve ecological properties of microbiomes. *Current Opinion in Systems Biology*, **8**, 73–80.

Needham, D.M., Sachdeva, R., & Fuhrman, J.A. (2017) Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *The ISME Journal*, **11**, 1614–1629.

Newman, M.E.J. (2002) Assortative Mixing in Networks. *Phys. Rev. Lett.*, **89**, 208701.

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*.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., & Stahl, D.A. (1986) Microbial Ecology and Evolution: A Ribosomal RNA Approach. *Annu. Rev. Microbiol.*, **40**, 337–365.

PARADA, A.E. & FUHRMAN, J.A. (2017) Marine archaeal dynamics and interactions with the microbial community over 5 years from surface to seafloor. *The ISME Journal*, **11**, 2510–2525.

PARADA, A.E., NEEDHAM, D.M., & FUHRMAN, J.A. (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, **18**, 1403–1414.

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.

PEOPLES, L.M., DONALDSON, S., OSUNTOKUN, O., XIA, Q., NELSON, A., BLANTON, J., ALLEN, E.E., CHURCH, M.J., & BARTLETT, D.H. (2018) Vertically distinct microbial communities in the Mariana and Kermadec trenches. *PLOS ONE*, **13**, 1–21.

PERNICE, M.C., GINER, C.R., LOGARES, R., PERERA-BEL, J., ACINAS, S.G., DUARTE, C.M., GASOL, J.M., & MASSANA, R. (2016) Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *The ISME Journal*, **10**, 945–958.

PICCARDI, P., VESSMAN, B., & MITRI, S. (2019) Toxicity drives facilitation between 4 bacterial species. *Proc Natl Acad Sci USA*, **116**, 15979.

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.

POISOT, T., CANARD, E., MOUILLOT, D., MOUQUET, N., & GRAVEL, D. (2012) The dissimilarity of species interaction networks. *Ecology Letters*, **15**, 1353–1361.

PRIM, R.C. (1957) Shortest connection networks and some generalizations. *The Bell System Technical Journal*, **36**, 1389–1401.

PRŽULJ, N., CORNEIL, D.G., & JURISICA, I. (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–3515.

PRZYTYCKA, T.M., SINGH, M., & SLONIM, D.K. (2010) Toward the dynamic interactome: it's about time. *Briefings in Bioinformatics*, **11**, 15–29.

PUFAHL, P.K. & HIATT, E.E. (2012) Oxygenation of the Earth's atmosphere–ocean system: A review of physical and chemical sedimentologic responses. *Marine and Petroleum Geology*, **32**, 1–20.

QAMAR, H., HUSSAIN, K., SONI, A., KHAN, A., HUSSAIN, T., & CHÉNAIS, B. (2021) Cyanobacteria as Natural Therapeutics and Pharmaceutical Potential: Role in Antitumor Activity and as Nanovectors. *Molecules*, **26**.

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.

RAHIMINEJAD, S., MAURYA, M.R., & SUBRAMANIAM, S. (2019) Topological and functional comparison of community detection algorithms in biological networks. *BMC Bioinformatics*, **20**, 212.

RASMUSSEN, B., FLETCHER, I.R., BROCKS, J.J., & KILBURN, M.R. (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature*, **455**, 1101–1104.

REVELLE, W. (2020) *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University.

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

RÖTTJERS, L. & FAUST, K. (2019) Can we predict keystones? *Nature Reviews Microbiology*, **17**, 193–193.

RÖTTJERS, L. & FAUST, K. (2020) manta: a Clustering Algorithm for Weighted Ecological Networks. *mSystems*, **5**.

RÖTTJERS, L., VANDEPUTTE, D., RAES, J., & FAUST, K. (2020) A framework for comparing microbial networks reveals core associations. *bioRxiv*, 2020.10.05.325860.

ROYO-LLONCH, M., SÁNCHEZ, P., RUIZ-GONZÁLEZ, C., SALAZAR, G., PEDRÓS-ALIÓ, C., LABADIE, K., PAOLI, L., CHAFFRON, S., EVEILLARD, D., KARSENTI, E., SUNAGAWA, S., WINCKER, P., KARP-BOSS, L., BOWLER, C., & ACINAS, S.G. (2020) Ecogenomics of key prokaryotes in the arctic ocean. *bioRxiv*, 2020.06.19.156794.

RUAN, Q., DUTTA, D., SCHWALBACH, M.S., STEELE, J.A., FUHRMAN, J.A., & SUN, F. (2006) Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*, **22**, 2532–2538.

RUIZ-GONZÁLEZ, C., LOGARES, R., SEBASTIÁN, M., MESTRE, M., RODRÍGUEZ-MARTÍNEZ, R., GALÍ, M., SALA, M.M., ACINAS, S.G., DUARTE, C.M., & GASOL, J.M. (2019) Higher contribution of globally rare bacterial taxa reflects environmental transitions across the surface ocean. *Molecular Ecology*, **28**, 1930–1945.

RUIZ-GONZÁLEZ, C., MESTRE, M., ESTRADA, M., SEBASTIÁN, M., SALAZAR, G., AGUSTÍ, S., MORENO-OSTOS, E., RECHE, I., ÁLVAREZ-SALGADO, X.A., MORÁN, X.A.G., DUARTE, C.M., SALA, M.M., & GASOL, J.M. (2020) Major imprint of surface plankton on deep ocean prokaryotic structure and activity. *Molecular Ecology*, **29**, 1820–1838.

SAGAN, L. (1967) On the origin of mitosing cells. *Journal of Theoretical Biology*, **14**, 225-IN6.

SALA, M.M., PETERS, F., GASOL, J.M., PEDRÓS-ALIÓ, C., MARRASÉ, C., & VAQUÉ, D. (2002) Seasonal and spatial variations in the nutrient limitation of bacterioplankton growth in the northwestern Mediterranean. *Aquatic Microbial Ecology*, **27**, 47–56.

SALAZAR, G., CORNEJO-CASTILLO, F.M., BENÍTEZ-BARRIOS, V., FRAILE-NUEZ, E., ÁLVAREZ-SALGADO, X.A., DUARTE, C.M., GASOL, J.M., & ACINAS, S.G. (2016) Global diversity and biogeography of deep-sea pelagic prokaryotes. *The ISME Journal*, **10**, 596–608.

SALAZAR, G., PAOLI, L., ALBERTI, A., HUERTA-CEPAS, J., RUSCHEWEYH, H.-J., CUENCA, M., FIELD, C.M., COELHO, L.P., CRUAUD, C., ENGELEN, S., GREGORY, A.C., LABADIE, K., MAREC, C., PELLETIER, E., ROYO-LLONCH, M., ROUX, S., SÁNCHEZ, P., UEHARA, H., ZAYED, A.A., ZELLER, G., CARMICHAEL, M., DIMIER, C., FERLAND, J., KANDELS, S., PICHERAL, M., PISAREV, S., POULAIN, J., ACINAS, S.G., BABIN, M., BORK, P., BOSS, E., BOWLER, C., COCHRANE, G., VARGAS, C. de, FOLLOWS, M., GORSKY, G., GRIMSLEY, N., GUIDI, L., HINGAMP, P., IUDICONE, D., JAILLON, O., KANDELS-LEWIS, S., KARP-BOSS, L., KARSENTI, E., NOT, F., OGATA, H., PESANT, S., POULTON, N., RAES, J., SARDET, C., SPEICH, S., STEMMANN, L., SULLIVAN, M.B., SUNAGAWA, S., WINCKER, P., ACINAS, S.G., BABIN, M., BORK, P., BOWLER, C., VARGAS, C. de, GUIDI, L., HINGAMP, P., IUDICONE, D., KARP-BOSS, L., KARSENTI, E., OGATA, H., PESANT, S., SPEICH, S., SULLIVAN, M.B., WINCKER, P., & SUNAGAWA, S. (2019) Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell*, **179**, 1068-1083.e21.

SANZ-SÁEZ, I. (2021) Contribution of marine heterotrophic cultured bacteria to microbial diversity and mercury detoxification. *Ph.D. thesis*.

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.

SCHIEBER, T.A., CARPI, L., DÍAZ-GUILERA, A., PARDALOS, P.M., MASOLLER, C., & RAVETTI, M.G. (2017) Quantification of network structural dissimilarities. *Nature Communications*, **8**, 13928.

SCHWARTZ, R.M. & DAYHOFF, M.O. (1978) Origins of Prokaryotes, Eukaryotes, Mitochondria, and Chloroplasts. *Science*, **199**, 395–403.

SCIUTO, K. & MORO, I. (2015) Cyanobacteria: the bright and dark sides of a charming group. *Biodiversity and Conservation*, **24**, 711–738.

SEBASTIÁN, M., ORTEGA-RETUERTA, E., GÓMEZ-CONSARNAU, L., ZAMANILLO, M., ÁLVAREZ, M., ARÍSTEGUI, J., & GASOL, J.M. (2021) Environmental and physical barriers drive the basin-wide spatial structuring of Mediterranean Sea and adjacent Eastern Atlantic Ocean prokaryotic communities. *Submitted*.

SEBASTIÁN, M., SÁNCHEZ, P., SALAZAR, G., ÁLVAREZ-SALGADO, X.A., RECHE, I., MORÁN, X.A.G., SALA, M.M., DUARTE, C.M., ACINAS, S.G., & GASOL, J.M. (2021) The quality of dissolved organic matter shapes the biogeography of the active bathypelagic microbiome. *bioRxiv*, 2021.05.14.444136.

SENDER, R., FUCHS, S., & MILO, R. (2016) Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, **14**, e1002533.

SEYMOUR, J.R., AMIN, S.A., RAINA, J.-B., & STOCKER, R. (2017) Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology*, **2**, 17065.

SHADE, A. & HANDELSMAN, J. (2012) Beyond the Venn diagram: the hunt for a core microbiome. *Environmental Microbiology*, **14**, 4–12.

SINCLAIR, L., OSMAN, O.A., BERTILSSON, S., & EILER, A. (2015) Microbial Community Composition and Diversity via 16S rRNA Gene Amplicons: Evaluating the Illumina Platform. *PLOS ONE*, **10**, e0116955.

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

SPORNS, O. (2018) Graph theory methods: applications in brain networks. *Dialogues Clin Neurosci*, **20**, 111–121.

STALEY, J.T. & KONOPKA, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual review of microbiology*, **39**, 321–346.

STAT, M., MORRIS, E., & GATES, R.D. (2008) Functional diversity in coral–dinoflagellate symbiosis. *Proc Natl Acad Sci USA*, **105**, 9256.

STEELE, J.A., COUNTWAY, P.D., XIA, L., VIGIL, P.D., BEMAN, J.M., KIM, D.Y., CHOW, C.-E.T., SACHDEVA, R., JONES, A.C., SCHWALBACH, M.S., ROSE, J.M., HEWSON, I., PATEL, A., SUN, F., CARON, D.A., & FUHRMAN, J.A. (2011) Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME Journal*, **5**, 1414–1425.

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.

SUNAGAWA, S., COELHO, L.P., CHAFFRON, S., KULTIMA, J.R., LABADIE, K., SALAZAR, G., DJAHANSCHIRI, B., ZELLER, G., MENDE, D.R., ALBERTI, A., CORNEJO-CASTILLO, F.M., COSTEA, P.I., CRUAUD, C., D'OVIDIO, F., ENGELEN, S., FERRERA, I., GASOL, J.M., GUIDI, L., HILDEBRAND, F., KOKOSZKA, F., LEPOIVRE, C., LIMA-MENDEZ, G., POULAIN, J., POULOS, B.T., ROYO-LLONCH, M., SARMENTO, H., VIEIRA-SILVA, S., DIMIER, C., PICHERAL, M., SEARSON, S., KANDELS-LEWIS, S., BOWLER, C., DE VARGAS, C., GORSKY, G., GRIMSLEY, N., HINGAMP, P., IUDICONE, D., JAILLON, O., NOT, F., OGATA, H., PESANT, S., SPEICH, S., STEMMANN, L., SULLIVAN, M.B., WEISSENBACH, J., WINCKER, P., KARSENTI, E., RAES, J., ACINAS, S.G., & BORK, P. (2015) Structure and function of the global ocean microbiome. *Science*, **348**, 1261359.

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., AYYAZ, 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.

TIKHONOV, M., LEACH, R.W., & WINGREEN, N.S. (2015) Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *The ISME Journal*, **9**, 68–80.

TRUNG, H.T., TOAN, N.T., VINH, T.V., DAT, H.T., THANG, D.C., HUNG, N.Q.V., & SATTAR, A. (2020) A comparative study on network alignment techniques. *Expert Systems with Applications*, **140**, 112883.

TSILIMIGRAS, M.C.B. & FODOR, A.A. (2016) Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, **26**, 330–335.

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.

VELLEND, M. (2020) *The theory of ecological communities (MPB-57)*. Princeton University Press.

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.

VIJAYAN, V., CRITCHLOW, D., & MILENKOVIĆ, T. (2017) Alignment of dynamic networks. *Bioinformatics*, **33**, i180–i189.
VIJAYAN, V., GU, S., KREBS, E.T., MENG, L., & MILENKOVIĆ, T. (2020) Pairwise Versus Multiple Global Network Alignment. *IEEE Access*, **8**, 41961–41974.
VIJAYAN, V. & MILENKOVIĆ, T. (2018) Aligning dynamic networks with DynaWAVE. *Bioinformatics*, **34**, 1795–1798.
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.
WALLIN, I.E. (1927) *Symbionticism and the Origin of Species*. Рипол Классик.
WANG, B., MEZLINI, A.M., DEMIR, F., FIUME, M., TU, Z., BRUDNO, M., HAIBE-KAINS, B., & GOLDENBERG, A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, **11**, 333.
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.
WANG, Y., YUAN, Y., MA, Y., & WANG, G. (2019) Time-Dependent Graphs: Definitions, Applications, and Algorithms. *Data Science and Engineering*, **4**, 352–366.
WARTON, D.I., BLANCHET, F.G., O'HARA, R.B., OVASKAINEN, O., TASKINEN, S., WALKER, S.C., & HUI, F.K.C. (2015) So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, **30**, 766–779.
WATTS, D.J. & STROGATZ, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.
WATTS, S.C., RITCHIE, S.C., INOUYE, M., & HOLT, K.E. (2019) FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics*, **35**, 1064–1066.
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.
WESTCOTT, S.L. & SCHLOSS, P.D. (2015) De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, **3**, e1487.
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.
XU, Z., WANG, M., WU, W., LI, Y., LIU, Q., HAN, Y., JIANG, Y., SHAO, H., MCMINN, A., & LIU, H. (2018) Vertical Distribution of Microbial Eukaryotes From Surface to the Hadal Zone of the Mariana Trench. *Frontiers in Microbiology*, **9**, 2023.
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.

YAVEROĞLU, Ö.N., MALOD-DOGNIN, N., DAVIS, D., LEVNAJIC, Z., JANJIC, V., KARAPANDZA, R., STOJMIROVIC, A., & PRŽULJ, N. (2014) Revealing the Hidden Language of Complex Networks. *Scientific Reports*, **4**, 4547.

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.

ZAHRA, Z., CHOO, D.H., LEE, H., & PARVEEN, A. (2020) Cyanobacteria: Review of Current Potentials and Applications. *Environments*, **7**.

ZHAO, D., SHEN, F., ZENG, J., HUANG, R., YU, Z., & WU, Q.L. (2016) Network analysis reveals seasonal variation of co-occurrence correlations between Cyanobacteria and other bacterioplankton. *Science of The Total Environment*, **573**, 817–825.

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.

ZHOU, J., DENG, Y., LUO, F., HE, Z., TU, Q., & ZHI, X. (2010) Functional Molecular Ecological Networks. *mBio*, **1**, e00169-10.

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.

Appendix

Long summary

There is a myriad of microorganisms on Earth contributing to global biogeochemical cycles, and their interactions are considered pivotal for ecosystem function. Previous studies have already determined relationships between a limited number of microorganisms. Yet, we still need to understand a large number of interactions to increase our knowledge of complex microbiomes. This is challenging because of the vast number of possible interactions. Thus, microbial interactions still remain barely known to date.

Networks are a great tool to handle the vast number of microorganisms and their connections, explore potential microbial interactions, and elucidate patterns of microbial ecosystems. The technological advances allowing omics-based censuses are helpful to infer microbial abundances over time or space. Such abundance-based data can be used to infer marine microbial association networks aggregated over time or spatial scales (single static networks). These association networks are a proxy to estimate ecological networks. Thus, microbial association networks are gaining popularity in marine microbial investigations. In these networks, nodes represent microorganisms and edges represent abundance-based associations between them. The associations provide microbial interaction hypotheses. Previous network-based investigations contributed to our understanding of marine microbial interactions. Much has been done in terms of network inference and there are numerous tools for that. However, there are diverse challenges beyond inference algorithms which need to be tackled to learn more from microbial networks.

This thesis locates at the intersection of network inference and network analysis. The presented methodology aims to support and advance marine microbial investigations by reducing noise and elucidating patterns in inferred association networks for subsequent biological down-stream analyses. The thesis comprises three main projects.

Subproject 1

The first challenge we tackled was the environmental influence on marine microbial association networks. An estimated association may represent a true ecological interaction or an indirect dependency due to environmental preference. Identifying indirect dependencies is a major challenge for inferring reliable microbial associations in the networks. Thus, such environmental effects need to be detected and excluded before downstream analysis.

This thesis's main contribution to marine microbial interactions studies is the development of the program EnDED (**En**vironmentally-**D**riven **E**dge **D**etection), a computational framework to identify environmentally-driven associations inside microbial association

networks, inferred from omics datasets. Identifying and removing environmentally-driven associations results in networks with edges more likely representing the most important ecological interactions between microorganisms. The methodology includes one new and three established techniques, which can be used individually or in the newly developed combination approach to exploit the combined power of multiple algorithms. Having all of them available in a single tool is what makes EnDED useful for analyzing microbial association networks. EnDED is a post-network inference program, i.e., it can be applied to networks inferred by different tools. The code is freely available with user-instructions and example code.

Moreover, since the evaluation of EnDED was hampered by the lack of gold-standard data, this thesis provided an adaptation of the generalized Lotka-Volterra (gLV) model to simulate relevant data on temporal microbial abundances influenced by environmental factors. We tested EnDED on the simulated data to systematically evaluate the different techniques performances with statistically sound metrics. The results showed that each individual method can detect moderate to high number of environmentally-driven edges (44-87%), and the combination approach resulted in retaining more associations but also more true interactions.

Furthermore, we tested if a simple threshold on the association strength could separate true from false associations as such a filtering step appeared to be common in marine microbial association network studies. Our results indicated that such threshold is not sufficient showing that we should apply better filtering strategies, e.g., future marine microbial association studies should include indirect-dependencies detection and their removal.

Subproject 2

The second challenge revolved around the usage of single static networks constructed from temporal data. Often, marine microbial association networks inferred from temporal data represent an aggregation, i.e., they contain permanent, temporary, or seasonal associations. The nature of an association needs to be disentangled. It is challenging to construct a temporal network consisting of a layer (a network itself) for each time point because of data sampling. Thus, we proposed a post network-construction approach to circumvent current dataset limitations by generating time-specific subnetworks from a single static network for each time point. These subnetworks represented the layers of the temporal network. Our methodology is a step forward for the study of marine microbial associations in time. Moreover, time-specific subnetworks allow to quantify an associations temporal recurrence. Thus, our methodology allows to characterize further marine microbial associations. Furthermore, quantifying the temporal recurrence of the associations may improve and shorten the list of interaction hypotheses for experimental testing.

We applied the methodology to a model marine microbial ecosystem at the Blanes Bay Microbial Observatory (BBMO) in the North-Western Mediterranean Sea (ten years of monthly sampling). Results indicated a one-year periodicity in the network topology. The temporal

architecture was not stochastic, but displayed a modest amount of recurrence over time, especially in winter. The marine microbial network appeared to collapse from winter to summer and reassemble from summer to winter each year during the ten years. Quantifying temporal recurrence indicated essential (core) associations within each season at the Blanes Bay microbial observatory. We found a much larger potential core in winter than in summer. Thus, the results indicated that marine microbial networks follow recurrent temporal dynamics, which need to be accounted to better understand the dynamics of the ocean microbiome.

Subproject 3

Similarly, to networks inferred from temporal data, networks that were inferred from spatial data represent an aggregation, i.e., they contain global and regional associations. In the third challenge, associations' biogeography, which is the distribution of associations, needs to be disentangled. We adapted the time-specific subnetwork approach to generate sample-specific subnetworks. These subnetworks allow to quantify an associations spatial recurrence allowing to determine global and regional associations. Thus, quantifying spatial recurrence characterized further marine microbial associations. Selecting high recurrent associations may improve and shorten the list of interaction hypotheses for experimental testing.

We applied the methodology to a dataset compilation covering six global-ocean regions from the surface (3 m) to the deep ocean (down to 4539 m). Thus, our methodology provided a step towards studying the marine microbial distribution in space via the horizontal (ocean regions) and vertical (water column) axes. We found the highest and the lowest fractions of global associations in the deep chlorophyll maximum (DCM) layer and the bathypelagic zone, respectively, whereas regional associations increased with depth. Our results indicated that associations have specific spatial distributions that are not just mirroring microbial distributions.

In addition, we employed local network metrics (based on graphlets) to cluster similar sample-specific subnetworks. Commonly, samples have been clustered on the basis of pre-defined grouping, e.g., ocean region and depth, or microbial compositions. However, pre-defined groups may introduce a bias, and the presence of two microorganisms is a necessary but not sufficient condition for the presence of a microbial interaction. Here, we introduced a new approach employing sample-specific subnetworks and local network metrics to cluster similar subnetworks. Our methodology was entirely focusing on network architecture, i.e., it is free from pre-defined groupings and does not take into account which specific microorganisms are present. We identified 36 clusters. Of these, 13 (36.1%) were dominated by surface subnetworks, and 11 (30.6%) by a deeper layer: 2 (5.6%) DCM, 5 (13.9%) mesopelagic, and 4 (11.1%) bathypelagic zone. Region-wise, we found 11 (30.6%) clusters exclusively or mainly containing subnetworks of the Mediterranean Sea, and only one (2.8%) cluster dominated by an ocean basin (North Atlantic Ocean).

Significance

This thesis focused on improving association networks for downstream analysis by disentangling environmental effects, temporal patterns, and the biogeography of associations. It did so by primarily employing graph-theoretic concepts, but it also drew from other fields: modeling microbial associations via an adjusted gLV model; or using information theoretic properties to determine indirect dependencies, and statistical measurements for evaluation. To reach accurate interaction hypotheses, it is important to determine, quantify, and remove environmentally-driven associations from marine microbial association networks. Thus, EnDED should be included in filter strategies. Moreover, our results underlined the need to study the dynamic nature of networks, in contrast to using single static networks aggregated over time or space. Our novel methodologies can be used by a wide array of researchers investigating networks and interactions in diverse microbiomes.

Keywords

- marine microbial association networks
- indirect dependencies
- environmentally-driven associations
- temporal network
- permanent versus temporary associations
- network collapse and reassembling
- sample-specific subnetworks
- biogeography of associations
- archaea, bacteria, and micro-eukaryotes

Resumen extenso

Hay una gran cantidad de microorganismos en la Tierra que contribuyen a los ciclos biogeoquímicos globales, y sus interacciones se consideran fundamentales para la función del ecosistema. Estudios previos ya han determinado relaciones entre un número limitado de microorganismos. Sin embargo, todavía necesitamos comprender una gran cantidad de interacciones para aumentar nuestro conocimiento de los microbiomas más complejos. Esto representa un gran desafío debido a la gran cantidad de posibles interacciones. Por lo tanto, las interacciones microbianas son aun poco conocidas.

Las redes representan una gran herramienta para analizar la gran cantidad de microorganismos y sus conexiones, explorar posibles interacciones y dilucidar patrones en ecosistemas microbianos. Los avances tecnológicos que hoy permiten censos microbianos basados en aproximaciones ómicas son útiles para inferir abundancias microbianas en el tiempo o el espacio. Estos datos basados en abundancias microbianas se pueden utilizar para inferir redes de asociación entre microbios marinos agregadas a lo largo del tiempo o a escalas espaciales (redes estáticas). Estas redes de asociación permiten estimar redes ecológicas. Por lo tanto, las redes de asociación microbiana están ganando popularidad en las investigaciones microbianas marinas. En estas redes, los nodos representan microorganismos y las conexiones representan asociaciones basadas en correlaciones de abundancia entre ellos. Las asociaciones proporcionan hipótesis de interacción microbiana. Investigaciones previas basadas en redes contribuyeron a nuestra comprensión de las interacciones microbianas marinas. Se ha avanzado significativamente en términos de inferencia de redes y existen numerosas herramientas para ello. Sin embargo, existen diversos desafíos más allá de los algoritmos de inferencia que deben abordarse para comprender más las redes microbianas.

Esta tesis se ubica en la intersección entre la inferencia de redes y el análisis de redes. La metodología presentada tiene como objetivo avanzar las investigaciones sobre interacciones microbianas marinas mediante la reducción del ruido en las inferencias de redes y elucidar patrones en redes de asociación permitiendo análisis biológicos posteriores. La tesis comprende tres proyectos principales.

Subproyecto 1

El primer desafío que abordamos fue la influencia ambiental en las redes de asociaciones de microbios marinos. Una asociación en las redes puede representar una interacción ecológica verdadera o una dependencia indirecta debido a la preferencia ambiental. La identificación de dependencias indirectas es un desafío importante para inferir asociaciones microbianas confiables en las redes. Por lo tanto, dichos efectos ambientales deben detectarse y excluirse antes del análisis posterior.

La principal contribución de esta tesis a los estudios de interacciones microbianas marinas es el desarrollo del programa EnDED (**En**vironmentally-**D**riven **E**dge **D**etection), un marco computacional para identificar asociaciones generadas por el medio ambiente en redes de asociaciones microbianas, inferidas a partir de datos ómicos. La identificación y eliminación de asociaciones generadas por el medio ambiente da como resultado redes con conexiones que probablemente representan las interacciones ecológicas más importantes entre microorganismos. La metodología incluye una técnica nueva y tres establecidas, que se pueden utilizar individualmente o combinadas para explotar el poder de múltiples algoritmos. Tener a todos ellos disponibles en una sola herramienta es lo que hace que EnDED sea útil para analizar redes de asociación microbiana. EnDED es un programa de inferencia *a posteriori*, es decir, se puede aplicar a redes inferidas por diferentes herramientas. El código está disponible gratuitamente con instrucciones para el usuario y código de ejemplo.

Dado que la evaluación de EnDED se vio obstaculizada por la falta de datos estándares, esta tesis proporcionó una adaptación del modelo generalizado de Lotka-Volterra (gLV) para simular datos relevantes sobre abundancias microbianas temporales influenciadas por factores ambientales. Probamos EnDED en los datos simulados para evaluar sistemáticamente el rendimiento de las diferentes técnicas. Los resultados mostraron que cada método individual puede detectar un número entre moderado y alto de conexiones generadas por el medio ambiente (44-87%). En cambio, el enfoque combinado retuvo más asociaciones pero también más interacciones verdaderas.

Además, probamos si el uso de un umbral de fuerza de la asociación por si sola podría separar las asociaciones verdaderas de las falsas, ya que tal paso de filtrado es común en los estudios de redes de asociaciones microbianas marinas. Nuestros resultados indicaron que dicho umbral no es suficiente, lo que demuestra que deberíamos aplicar mejores estrategias de filtrado, por ejemplo, los futuros estudios de asociación de microbios marinos deberían incluir la detección de dependencias indirectas y su eliminación.

Subproyecto 2

El segundo desafío de esta tesis gira en torno al uso de redes estáticas únicas construidas a partir de datos temporales. A menudo, las redes de asociación de microbios marinos inferidas a partir de datos temporales representan una agregación, es decir, contienen asociaciones permanentes, temporales o estacionales. Es necesario determinar la naturaleza de cada asociación. Construir una red temporal que consta de una capa (una red en sí misma) para cada punto de tiempo es un desafío debido al muestreo de datos. Por lo tanto, propusimos un enfoque posterior a la construcción de la red para eludir las limitaciones actuales del conjunto de datos mediante la generación de subredes específicas en el tiempo a partir de una única red estática. Estas subredes representaron las capas de la red temporal. Nuestra metodología es un paso adelante para el estudio de las asociaciones microbianas marinas en el tiempo. Además, las subredes específicas

en el tiempo permiten cuantificar la recurrencia temporal de una asociación. Por lo tanto, nuestra metodología permite caracterizar más asociaciones microbianas marinas. Además, la cuantificación de la recurrencia temporal de las asociaciones puede mejorar y acortar la lista de hipótesis de interacción para las pruebas experimentales.

Aplicamos la metodología a un modelo de ecosistema microbiano marino en el Observatorio Microbiano de la Bahía de Blanes (BBMO) en el Mar Mediterráneo Noroccidental (diez años de muestreo mensual). Los resultados indicaron una periodicidad de un año en la topología de la red. La arquitectura temporal no fue estocástica, pero mostró una modesta cantidad de recurrencia en el tiempo, especialmente en invierno. La red microbiana marina pareció colapsar de invierno a verano y volver a formarse de verano a invierno cada año durante los diez años. La cuantificación de la recurrencia temporal indicó asociaciones esenciales (centrales) dentro de cada temporada en el observatorio microbiano de la bahía de Blanes. Encontramos un microbioma núcleo potencial mucho mayor en invierno que en verano. Por lo tanto, los resultados indicaron que las redes microbianas marinas siguen dinámicas temporales recurrentes, que deben tenerse en cuenta para comprender mejor la dinámica del microbioma oceánico.

Subproyecto 3

De manera similar a las redes inferidas a partir de datos temporales, las redes inferidas a partir de datos espaciales representan una agregación, es decir, contienen asociaciones globales y regionales. En el tercer desafío, se busca identificar la biogeografía de las asociaciones, que es la distribución de las mismas en el espacio. Adaptamos el enfoque de subred de tiempo específico para generar subredes espaciales específicas de muestra. Estas subredes permiten cuantificar la recurrencia espacial de las asociaciones, permitiendo identificar asociaciones globales y regionales. La selección de asociaciones de alta recurrencia puede mejorar y acortar la lista de hipótesis de interacción para pruebas experimentales.

Aplicamos la metodología a una compilación de conjuntos de datos que cubren seis regiones oceánicas globales desde la superficie (3 m) hasta las profundidades del océano (hasta 4539 m). Por lo tanto, nuestra metodología significa un paso adelante hacia el estudio de la distribución microbiana marina en el espacio a través de los ejes horizontal (regiones oceánicas) y vertical (columna de agua). Encontramos las fracciones más alta y más baja de asociaciones globales en la capa máxima de clorofila profunda (DCM) y la zona batipelágica, respectivamente, mientras que las asociaciones regionales aumentaron con la profundidad. Nuestros resultados indicaron que las asociaciones tienen distribuciones espaciales específicas que no solo reflejan las distribuciones taxonómicas microbianas.

Además, empleamos métricas de red local (basadas en graphlets) para agrupar subredes similares específicas de muestras. Por lo general, las muestras se han agrupado sobre la base de agrupaciones predefinidas, por ejemplo, región y profundidad del océano, o composiciones microbianas. Sin embargo, los grupos predefinidos pueden introducir un sesgo, y la presencia de

dos microorganismos es una condición necesaria pero no suficiente para la presencia de una interacción microbiana. Aquí, presentamos un nuevo enfoque que emplea subredes específicas de muestra y métricas de red local para agrupar subredes similares. Nuestra metodología se centró completamente en la arquitectura de red, es decir, está libre de agrupaciones predefinidas y no tiene en cuenta qué microorganismos específicos están presentes. Identificamos 36 conglomerados. De estos, 13 (36,1%) estaban dominados por subredes de superficie y 11 (30,6%) por una capa más profunda: 2 (5,6%) DCM, 5 (13,9%) mesopelágicas y 4 (11,1%) zona batipelágica. Por región, encontramos 11 (30,6%) conglomerados que contienen exclusiva o principalmente subredes del Mar Mediterráneo, y solo un conglomerado (2,8%) dominado por una cuenca oceánica (Océano Atlántico Norte).

Significado

Esta tesis se centró en mejorar las redes de asociación para el análisis posterior al identificar los efectos ambientales, los patrones temporales y la biogeografía de las asociaciones. Lo hizo empleando principalmente conceptos de teoría de gráficos, pero también se basó en otros campos: modelado de asociaciones microbianas a través de un modelo de gLV ajustado; o usar propiedades teóricas de la información para determinar dependencias indirectas y mediciones estadísticas para evaluación. Para llegar a hipótesis de interacción precisas, es importante determinar, cuantificar y eliminar las asociaciones generadas por el medio ambiente en las redes de asociaciones microbianas marinas. Por tanto, EnDED debería incluirse en las estrategias de filtrado. Además, nuestros resultados subrayaron la necesidad de estudiar la naturaleza dinámica de las redes, en contraste con el uso de redes estáticas únicas agregadas en el tiempo o el espacio. Nuestras nuevas metodologías pueden ser utilizadas por una amplia gama de investigadores que investigan redes e interacciones en diversos microbiomas.

Palabras clave

- redes de asociaciones de microbios marinos
- dependencias indirectas
- asociaciones impulsadas por el medio ambiente
- red temporal
- asociaciones permanentes versus temporales
- colapso y reensamblaje de la red
- subredes específicas de muestra
- biogeografía de asociaciones
- arqueas, bacterias y microeucariotas

Resum extens

Hi ha una infinitat de microorganismes a la Terra que contribueixen als cicles biogeoquímics mundials i les seves interaccions es consideren fonamentals pel funcionament dels ecosistemes. Estudis previs ja han determinat les relacions entre un nombre limitat de microorganismes. Tot i això, encara hem d'entendre un gran nombre d'interaccions per augmentar el nostre coneixement dels microbiomes complexos. Això és un repte a causa del gran nombre d'interaccions possibles. Per això, les interaccions microbianes encara són poc conegudes fins ara.

Les xarxes són una gran eina per tractar el gran nombre de microorganismes i les seves connexions, explorar interaccions microbianes potencials i dilucidar patrons d'ecosistemes microbians. Els avenços tecnològics que permeten censos basats en òmics són útils per inferir abundàncies microbianes al llarg del temps o de l'espai. Aquestes dades basades en l'abundància es poden utilitzar per inferir xarxes d'associació microbiana marina agregades al llarg del temps o en escales espacials (xarxes estàtiques individuals). Aquestes xarxes d'associació són una referencia per estimar les xarxes ecològiques. Així, les xarxes d'associació microbiana guanyen popularitat en investigacions microbianes marines. En aquestes xarxes, els nodes representen microorganismes i les vores representen associacions basades en l'abundància entre ells. Les associacions proporcionen hipòtesis d'interacció microbiana. Les investigacions prèvies basades en xarxes han contribuït a la nostra comprensió de les interaccions microbianes marines. S'ha fet molt en termes d'inferència de xarxes i hi ha nombroses eines per fer-ho. No obstant això, hi ha diversos reptes més enllà dels algorismes d'inferència que cal abordar per obtenir més informació de les xarxes microbianes.

Aquesta tesi es sitúa a la intersecció de la inferència de xarxes i l'anàlisi de la xarxes. La metodologia presentada té com a objectiu donar suport i avançar en investigacions microbianes marines reduïnt el soroll i dilucidant patrons en xarxes d'associació inferides per a posteriors anàlisis biològiques. La tesi comprèn tres projectes principals.

Subprojecte 1

El primer repte que vam abordar va ser la influència ambiental en les xarxes d'associació microbiana marina. Una associació estimada pot representar una interacció ecològica real o una dependència indirecta a causa de la preferència ambiental. Identificar dependències indirectes és un repte important per inferir associacions microbianes fiables a les xarxes. Per tant, aquests efectes ambientals han de ser detectats i exclosos abans de l'anàlisi posterior.

La principal contribució d'aquesta tesi als estudis d'interaccions microbianes marines és el desenvolupament del programa EnDED (**En**vironmentally-**D**riven **E**dge **D**etection), un marc computacional per identificar associacions impulsades pel medi ambient dins de xarxes

d'associació microbiana, inferides a partir de conjunts de dades òmics. La identificació i eliminació d'associacions impulsades pel medi ambient dóna lloc a xarxes amb arestes que probablement representen les interaccions ecològiques més importants entre microorganismes. La metodologia inclou una tècnica nova i tres establertes, que es poden utilitzar individualment o en un nou enfocament combinat, desenvolupat per explotar la potència combinada de múltiples algorismes. Tenir-los tots disponibles en una sola eina és el que fa que EnDED sigui útil per analitzar xarxes d'associació microbiana. EnDED és un programa d'inferència post-xarxa, és a dir, que es pot aplicar a xarxes inferides per diferents eines. El codi està disponible gratuïtament amb instruccions d'usuari i codi d'exemple.

A més, atès que l'avaluació d'EnDED es va veure obstaculitzada per la manca de dades estàndard, aquesta tesi va proporcionar una adaptació del model generalitzat de Lotka-Volterra (gLV) per simular dades rellevants sobre les abundàncies microbianes temporals influïdes per factors ambientals. Hem provat EnDED amb les dades simulades per avaluar sistemàticament les diferents tècniques de rendiment amb mètriques estadísticament sòlides. Els resultats van mostrar que cada mètode individual pot detectar un nombre moderat a elevat d'arestes ambientals (44-87%), i l'enfocament combinat va donar lloc a mantenir més associacions però també interaccions més genuines.

A més, vam provar si un llindar simple sobre la força de l'associació podia separar-se de les associacions falses, ja que aquest pas de filtratge semblava ser comú en els estudis de xarxes d'associació microbiana marina. Els nostres resultats van indicar que aquest llindar no és suficient per demostrar que hauríem d'aplicar millors estratègies de filtratge, per exemple, els futurs estudis d'associació microbiana marina haurien d'incloure la detecció de dependències indirectes i la seva eliminació.

Subprojecte 2

El segon repte va girar al voltant de l'ús de xarxes estàtiques individuals construïdes a partir de dades temporals. Sovint, les xarxes d'associació microbiana marina inferides a partir de dades temporals representen una agregació, és a dir, contenen associacions permanents, temporals o estacionals. Cal desentranyar la naturalesa d'una associació. És difícil crear una xarxa temporal que consisteixi en una capa (una xarxa pròpia) per a cada punt temporal a causa del mostreig de dades. Per tant, vam proposar un enfocament post-construcció de xarxes per eludir les limitacions actuals dels conjunts de dades mitjançant la generació de subxarxes específiques en el temps a partir d'una única xarxa estàtica per a cada punt de temps. Aquestes subxarxes representaven les capes de la xarxa temporal. La nostra metodologia és un pas endavant per a l'estudi de les associacions microbianes marines en el temps. A més, les subxarxes específiques en el temps permeten quantificar una recurrència temporal de les associacions. Per tant, la nostra metodologia permet caracteritzar altres associacions microbianes marines. A més, quantificar la recurrència

temporal de les associacions pot millorar i escurçar la llista d'hipòtesis d'interacció per a proves experimentals.

Vam aplicar la metodologia a un model d'ecosistema microbià marí a l'Observatori Microbià de la Badia de Blanes (BBMO) al mar Mediterrani nord-occidental (deu anys de mostreig mensual). Els resultats van indicar una periodicitat d'un any en la topologia de la xarxa. L'arquitectura temporal no era estocàstica, sino que presentava una petita quantitat de recurrència al llarg del temps, especialment a l'hivern. La xarxa microbiana marina semblava col·lapsar-se d'hivern a estiu i es tornava a muntar d'estiu a hivern cada any durant els deu anys. Quantificar la recurrència temporal va indicar associacions essencials (bàsiques) dins de cada temporada a l'observatori microbià de la badia de Blanes. Hem trobat un nucli potencial molt més gran a l'hivern que a l'estiu. Així, els resultats van indicar que les xarxes microbianes marines segueixen dinàmiques temporals recurrents, que cal tenir en compte per comprendre millor la dinàmica del microbioma oceànic.

Subprojecte 3

De manera similar a les xarxes inferides a partir de dades temporals, les xarxes inferides a partir de dades espacials representen una agregació, és a dir, contenen associacions globals i regionals. En el tercer desafiament, cal desentrellaçar la biogeografia de les associacions, que és la distribució de les associacions. Vam adaptar l'enfocament de subxarxes específiques del temps per generar subxarxes específiques de mostra. Aquestes subxarxes permeten quantificar una recurrència espacial de les associacions permetent determinades associacions globals i regionals. Per tant, quantificar la recurrència espacial va caracteritzar altres associacions microbianes marines. La selecció d'associacions recurrents altes pot millorar i escurçar la llista d'hipòtesis d'interacció per a proves experimentals.

Vam aplicar la metodologia a una recopilació de dades que cobreix sis regions oceàniques globals des de la superfície (3 m) fins a l'oceà profund (fins a 4539 m). Per tant, la nostra metodologia va proporcionar un pas cap a l'estudi de la distribució microbiana marina a l'espai a través dels eixos horitzontal (regions oceàniques) i vertical (columna d'aigua). Hem trobat les fraccions més altes i les més baixes d'associacions mundials a la capa clorofil·la profunda màxima i a la zona batipelàgica, respectivament, mentre que les associacions regionals augmentaven amb la profunditat. Els nostres resultats van indicar que les associacions tenen distribucions espacials específiques que no són només rèpliques de distribucions microbianes.

A més, hem emprat mètriques de xarxa local (basades en gràfics) per agrupar subxarxes similars específiques de mostra. Normalment, les mostres s'han agrupat sobre la base d'un agrupament predefinit, per exemple, la regió i la profunditat de l'oceà o les composicions microbianes. No obstant això, els grups predefinits poden introduir un biaix i la presència de dos microorganismes és una condició necessària però no suficient per a la presència d'una interacció

microbiana. Aquí hem introduït un nou enfocament que utilitza subxarxes específiques de mostra i mètriques de xarxa local per agrupar subxarxes similars. La nostra metodologia es va centrar completament en l'arquitectura de xarxa, és a dir, està lliure d'agrupacions predefinides i no té en compte quins microorganismes específics hi són presents. Vam identificar 36 clústers. D'aquests, 13 (36,1%) estaven dominats per subxarxes superficials i 11 (30,6%) per una capa més profunda: 2 (5,6%) de DCM, 5 (13,9%) mesopelàgica i 4 (11,1%) zona batipelàgica. Per regions, vam trobar 11 (30,6%) clústers que contenien exclusivament o principalment subxarxes del mar Mediterrani, i només un (2,8%) dominat per una conca oceànica (oceà Atlàntic Nord).

Importància

Aquesta tesi s'ha centrat en la millora de les xarxes d'associacions per a l'anàlisi posterior, desfent els efectes ambientals, els patrons temporals i la biogeografia de les associacions. Això s'ha fet emprant principalment conceptes de teoría de grafs, però també s'ha basat en altres camps: modelatje d'associacions microbianes mitjançant un model de gLV ajustat; o l'ús propietats teòriques de la informació per determinar dependències indirectes i mesures estadístiques per a la seva avaluació. Per arribar a hipòtesis d'interacció precises, és important determinar, quantificar i eliminar associacions impulsades pel medi ambient de les xarxes d'associació microbiana marina. Per tant, EnDED s'hauria d'incloure a les estratègies de filtratge. A més, els nostres resultats van subratllar la necessitat d'estudiar la naturalesa dinàmica de les xarxes, en contrast amb l'ús de xarxes estàtiques individuals agregades al llarg del temps o l'espai. Les nostres noves metodologies poden ser utilitzades per una àmplia gamma d'investigadors que investiguen xarxes i interaccions en diversos microbiomes.

Paraules clau

- xarxes d'associació microbiana marina
- dependències indirectes
- associacions impulsades pel medi ambient
- xarxa temporal
- associacions permanents versus temporals
- collapse i remuntatge de la xarxa
- subxarxes específiques de mostra
- biogeografia d'associacions
- arqueus, bacteris i microeucariotes

This thesis was brought to you thanks to the support and encouragement of my family and friends.