ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Microbial communities through the lens of high throughput sequencing, data integration and metabolic networks analysis

Haris Zafeiropoulos

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Science (PhD) in Biology

Promotors:
Prof. Emmanouil Ladoukakis
Dr Evangelos Pafilis
Dr Christoforos Nikolaou

Academic year 2021 – 2022

Members of the examination committee
&
reading committee

Prof. Emmanouil Ladoukakis
Univeristy of Crete
Biology Department

Dr. Evangelos Pafilis
Hellenic Centre for Marine Research
Institute of Marine Biology, Biotechnology and Aquaculture

Dr. Christoforos Nikolaou
Biomedical Sciences Research Center “Alexander Fleming”
Institute of Bioinnovation

Prof. Konstadia (Dina) Lika
Univeristy of Crete
Biology Department

Prof. Panagiotis Sarris
University of Crete
Department of Biology

Dr. Jens Carlsson
University College Dublin
School of Biology and Environmental Science/Earth Institute

Prof. Karoline Faust
KU Leuven
Department of Microbiology and Immunology, Rega Institute

Preface

"The problem is to construct a third view, one that sees the entire world neither as an indissoluble whole nor with the equally incorrect, but currently dominant, view that at every level the world is made up of bits and pieces that can be isolated and that have properties that can be studied in isolation. Both ideologies [...] prevent a rich understanding of nature and prevent us from solving the problems to which science is supposed to apply itself."
- **Lewontin**, Biology as Ideology

" Thus at every step we are reminded that we by no means rule over nature like a conqueror over a foreign people, like someone standing outside nature - but that we, with flesh, blood and brain, belong to nature, and exist in its midst, and that all our mastery of it consists in the fact that we have the advantage over all other creatures of being able to learn its laws and apply them correctly."
- **Friedrich Engels**, Dialectics of Nature

It is quite common at this point to thank those who were there for you, who helped with your work and your mental health. It makes perfect sense to put aside for a while all the struggle and the things that make us feel bad and say some nice words that can help us keep going on. As this is where I can share some of my thoughts though, if I did so, then this it would not be me.

Alan's Moore character *V* mentioned with great style something about his fictional England: *"And the truth is, there is something terribly wrong with this country, isn't there"*. I find this quote describe , not just about Greece though, but in world-wide extent.

“Knowledge, like air, is vital to life. Like air, no one should be denied it.” “Everybody is special. Everybody. Everybody is a hero, a lover, a fool, a villain. Everybody. Everybody has their story to tell.”

covid made things crystal clear.

The reason I fist started this PhD and the reason I am on this job is to contribute in changing this.

To me, this thesis is the sequel of my MSc. This work would never have started if it was not for: Along with our tutors Alexandros Kanterakis, Spiros Chavlis, Christoforos Nikolaou, Pavlos Pavlidis, Michalis Tsagris (typical example of a good guy from Greece' *"holly grounds"*) all of them, each one on his/her own way, pointed out aspects of our science I would have never think on my own. But above all, in a dark

I have to admit that there is always a sly smile on my face every time I think of them and I consider myself privileged having most of them in my corner.

At this point I would like to thank my promotors: Prof. Manolis Ladoukakis; a university teacher who said *"relax; I will join your committee"* to a sweaty guy who invaded his office crying for help at the last moment; since then he has always been there when it was needed and his guidance has been .

Dr. Evangelos Pafilis with whom we started working together back in 2017 and thanks to him I had the opportunity to

Dr. Christoforos Nikolaou who has been an influence to me since I first came in Crete; it is not just his morals and perspective on science, but being a Liverpool fan and bibliophagist (I googled and this word does exists) I found someone to make me talk about literature again.

In this context, I would not like to thank but to pay my respects to those that endure and insist (not) being *common people*.

Tsocha, louloudi.. Pascha Kliou

Psilos.. best PhD body ever. We ll join microbiology and biogeogeography someday

Savvas no way I would finish this without him! After this thesis, I can go diving I am sure.

Tzortzia, Chlapatsa ...

Vouno tsigaraki

Stelio, Antoni

Tweens power

Mom, dad Granny

Afous

Elefantas - xazi - ioannaki –> morakia

Haris Zafeiropoulos

Contents

Abstract

Microbial communities are a cornerstone for most ecosystem types. To elucidate the mechanisms governing such assemblages, it is fundamental to identify the taxa present (*who*) and the processes that occur (*what*) in the various environments (*where*). Thanks to a series of technological breakthroughs vast amounts of information/data from all the various levels of the biological organization have been accumulated over the last decades. In this context, microbial ecology studies are now relying on bioinformatics methods and analyses. Therefore, a great number of challenges both from the biologist- and the computer scientist point-of-view have arisen; one among the most emerging ones being: *"what shall we do with all these pieces of information?"*. The paradigm of Systems Biology addresses this challenge by moving from reductionism to more holistic approaches attempting to interpret how the properties of a system emerge.

Aim of this PhD was to enhance microbiome data analyses by developing software addressing on-going computational challenges on the study of microbial communities. On top of that, to exploit state-of-the-art methods to identify taxa, functions and microbial interactions in assemblages of various aquatic environments. To this end, a number of publicly available data-sets were used while a swamp from the Karpathos island (Greece), was chosen as a study case for the described framework.

Environmental DNA and metabarcoding have been widely used to estimate the biodiversity (the *who*) and the structure of communities. Vast amount of sequencing data targeting certain marker genes depending the taxonomic group of interest become available thanks to High Throughput Sequencing technologies. However, the bioinformatics analysis of such data require multiple steps and parameter settings. Software workflow-oriented tools along with computing infrastructures ease this need to a great extent and PEMA was developed to this end (Chapter **??**). However, eDNA metabarcoding has limitations too. Cytochrome c oxidase subunit I (COI) marker gene is a commonly used marker gene, especially in studies targeting eukaryotic taxa. It is well known that in COI studies a great number of the derived OTUs/ASVs get no taxonomic hits. The presence of non-eukaryotic taxa with their simultaneous absence from the most commonly-used reference databases justify this phenomenon to a great extent. DARN makes use of a COI-oriented tree of life to provide further insight to such known unknown sequences (Chapter **??**).

Shotgun metagenomics provide further information regarding the processes that occur in a community (the *what*). Sediment and microbial mat samples as well as microbial aggregates from a hypersaline swamp in Tristomo bay (Karpathos, Greece) were analyzed. Both amplicon (16S rRNA) and shotgun sequencing data were used to characterize the microbial structure of the communities and environmental parameters (e.g. salinity, oxygen concentration, granulometric composition) were measured at the sampling sites. Key functions supporting life in such environments were identified and metagenome-assembled genomes (MAGs) of major species found were built (Chapter **??**).

Amplicon and shotgun metagenomics approaches along with the rest of the omics technologies, have led to vast amount of data and metadata, recording the *who*, the *what* and the *where*. To enable optimal accessibility and usage of this information, a great number of databases, ontologies as well as community-standards have been developed. By exploiting data integration techniques to bring such bits of information together, as well as text mining methods to retrieve knowledge "hidden" among the billions of text lines in already published literature, the PREGO knowledge-base generates thousands of *what* - *where* - *who* potential associations (Section **??**).

The driving question though is *how* the different microbial taxa ascertain their endurance as part of a community. Metabolic interactions among the various taxa play a decisive role for the composition of such assemblages. Genome-scale metabolic networks (GEMs) enable the inference of such interactions. Random sampling on the flux space of such metabolic models, provides a representation of the flux values a model can get under various conditions. However, flux sampling is challenging from a computational point of view. To address such challenges, a Python library called dingo was developed using a Multiphase Monte Carlo Sampling algorithm (Chapter **??**). GEMs were built using the MAGs retrieved from the Tristomo swamp and metabolic interactions between them and their environment were investigated.

Similar to microbial communities, bioinformatics methods tend to build assemblages while "living" on your own is quite rare. The methods developed during this PhD project combined with state-of-the-art methods anticipate to build a framework that enables moving from the community to the species level and then back again to the one of the community.

Περίληψη

Και στα ελληνικά δουλευει σωστά

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

COI	Cytochrome c oxidase subunit I
ITS	Internal Transcribed Spacer
NGS	Next Generation Sequencing
eDNA	environmental DeoxyriboNucleic Acid
OTU	Operational Taxonomic Unit
ASV	Amplicon Sequence Variant
HPC	High Performance Computing
MCMC	Markov Chain Monte Carlo
MMCS	Multiphase Monte Carlo Sampling
PREGO	PRocess Environment OrGanism
PEMA	Pipeline for Environmental DNA Metabarcoding Analysis
DARN	Dark mAtteR iNvestigator
PSRF	Potential Scale Reduction Factor
ESS	Effective Sample Size

Symbols

$\mathbb{R}$	Set of real numbers
$\mathcal{O}$	Algorithm complexity
$\widetilde{\mathcal{O}}$	Algorithm complexity ignoring polylogarithmic factors
$\mathbb{E}$	Expected Value operator
$\mathcal{U}$	utility function....
ℓ	a ray
P	a polytope
τ	a trajectory
∂P	boundary of the P polytope
v	a flux vector
v_i	flux value of the reaction i
W	walk length
ρ	number of reflections

Chapter 1

Introduction

1.1 Microbial ecology in the 'omics era

1.1.1 Microbial communities: composition , functions & intections

Microbes are considered to be omnipresent in the various ecosystems on Earth [1]. It was only until recently, 2019, that scientists discovered for the first time a place on Earth where no microbial forms of life are present [2]. Extremely low pH, high salt and high temperature had to be at the same place at the same time to stop microbes. However, microbes are not just abundant but exceedingly variant too. Locey and Lennon using a unified scaling law and a lognormal model of biodiversity, estimated microbial diversity at about 1 trillion species [3]. However, despite the extensive studies of the scientific community, less than 1% of the microbial species on Earth have been identified [4].

Microbes are distinguished by multiple properties. Based on their morphology microbes can be spherical (cocci), rod-shaped (bacilli), arc-shaped (vibrio), and spiral (spirochete) [5]. Based on their metabolic characteristics, microbes are further distinguished. More specifically, according to their *energy source*, a microbe can either oxidate inorganic compounds (**chemotrophs**) or sunlight (**phototrophs**). Similarly, microbes can use CO_2 (**autotrophs**) as their *carbon source*, or organic compounds (**heterotrophs**) or both (**mixotrophs**). Finally, based on their *electron source* microbes are distinguished bewtween those using inorgarnic compounds (**lithotrophs**) and those using organic compounds (**organotrophs**) [6]. Microbial taxa combine combinining alternatives of the aforementioned categories shape a range of microbial profile of all the possible combinations; for example **chemolithoautotrophic** bacteria, e.g. nitrifying and sulfur-oxidizing bacteria, as well as **photoautotrophic** bacteria, e.g. purple bacteria and Green sulfur bacteria. Finally, microbia taxa can also be disinguished by their various ecological distributions and activities, and by their distinct genomic structure, expression, and evolution [5].

However, it is not only the number of microbial taxa and their massive biomass that make the study of microbial communities essential; it is mostly their functional potentials. Life on Earth would not be as we know it, if existed at all, if it was not for the microbes and their long contribution on ensuring life-supporting conditions. Nevertheless, these are the *biological machines responsible for planetary biogeochemical cycles* [1]; meaning that biogeochemical cycling to a global extent is powered by the metabolic processes of the microbial taxa [7]. In Figure 1.1 the contibution of microbial communities in the cycle of CO_2 is shown.

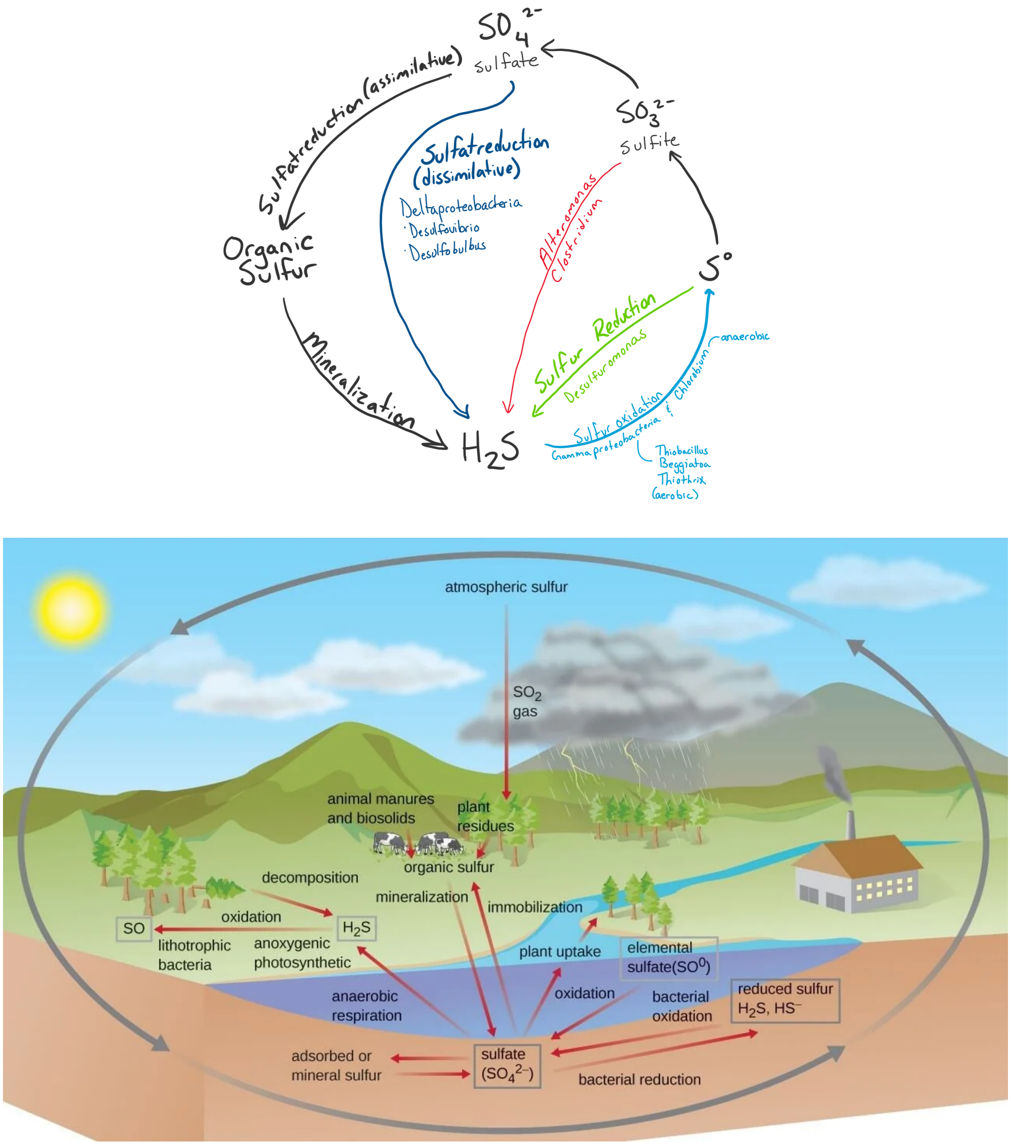

FIGURE 1.1: The cycle of sulfur (S) (up) and the contribution of microbial communities on it (down, image source: OpenStax).

The biological fluxes of most of the major elements (i.e., carbon, hydrogen, oxygen, nitrogen and sulfur) required for any biological macromolecule, are driven largely by microbially catalyzed, thermodynamically constrained redox reactions [1]. Phosphorus the last of the 6 fundamental elements for life, is also included in the metabolic pathways catalyzed by microbes. Thus, microbial communities consist of hundreds or even thousands of metabolically diverse strains and species [8], and their functions and determine the fitness of most organisms on Earth. In case of human health, specific microbial enzymatic pathways and molecules necessary for health promotion have been well known. Some of these "beneficial factors" are already known for probiotics and species in the human microbiome [9].

Microbial ecology studies the interactions:

- between microbial taxa and their environment
- among the various microbial taxa present in a community, and
- between microbial taxa and their host [4]

Microbial ecologists also investigate the role of microbial taxa in biogeochemical cycles [1] and their interaction with anthropogenic effects e.g. pollution and climate change [10].

Even though HTS has allowed a massive extension of our knowledge in specific enzymatic reactions that regulate these pathways the rules that determine the assembly, function, and evolution of these microbial communities remain unclear. Thus, both in case of environmental and human the underlying mechanisms for how microbial assemblages work and affect their environment, remain to be discovered. Understanding the underlying governing principles is central to microbial ecology [11] and crucial for designing microbial consortia for biotechnological [12] or medical applications [13].

Studies such as the one of Louca et al. have opened new frontiers in our understanding on microbial assemblages. After building metabolic functional groups and assigning more than 30,000 marine species to these groups, Louca et al. showed that the distribution of these functional groups were influenced by environmental conditions to a great extent, shaping *metabolic niches*. At the same time though, the taxonomic composition within individual functional groups were not affected by such environmental condintions [7].

Moreover, to elucidate how these assemblages work the biotic interactions have to be considered too. Microbial interactions play a fundamental role in deciphering the underlying mechanisms that govern ecosystem functioning [14, 15]. Microbes secrete costly metabolites (called **byproducts**) to their environment, which other microbes can absorb and exploit [16]. By exchanging metabolic products, mostly as there are also other ways of interactions e.g. quorum sensing, microbial taxa establish various interactions.

The interaction between two taxa can either be nutral or positive / negative. In case of a positive interaction, there is a case where both taxa benefit one from another. This *win-win* relationship is called **mutualism** (or "cooperation") and it can be a result of *cross-feeding*, in which two species exchange metabolic products [15]. Such is the case in biofilms where multiple bacterial taxa are working together building a structure that provides them antibiotic resistance [17]. There is also the case where only one of the two taxa benefits without helping or harming the other; this interaction is called **commensalism** [15]. For example, *Nitrosomonas* oxidize ammonia (NH_3) into nitrite (NO_2^-), so *Nitrobacter* can use it to obtain energy and oxidize it into nitrate (NO_3^-) [18]. Such interactions are quite common in microbial communities.

In case of a negative interaction, can harm each other either way (**compe-tition**). That is the case between *Listeria monocytogenes* and *Lactococcus lactis* in the study

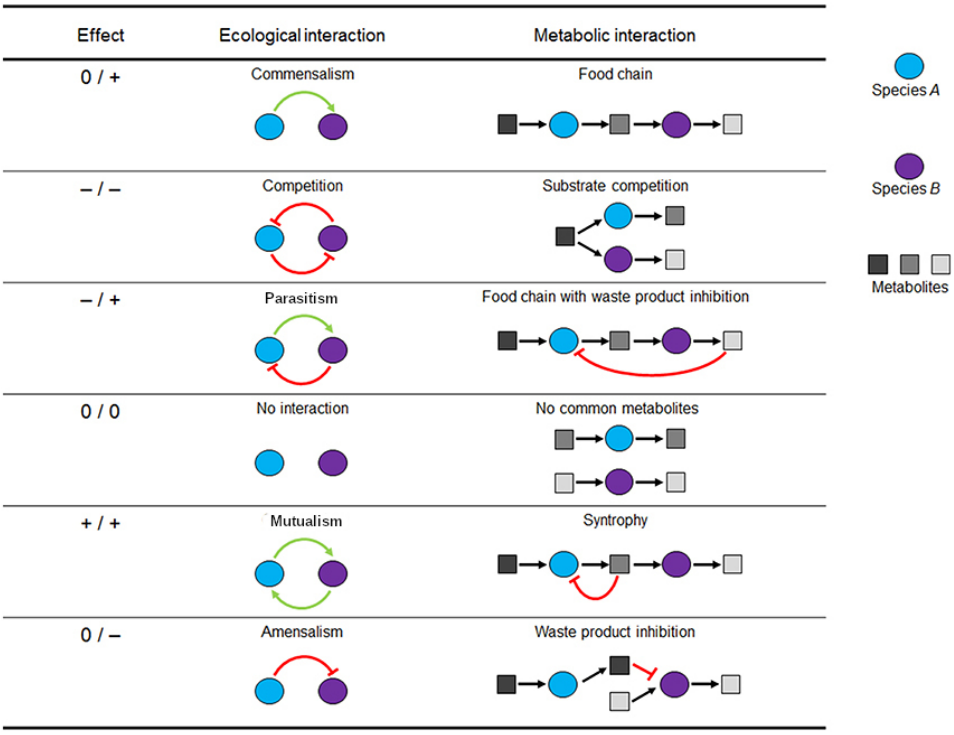

FIGURE 1.2: Microbial interaction types along with their corresponding metabolic ones. Due to certain metabolic interactions, two taxa may have a positive, a negative or a nutral effect one another. Figure based on [19]

of Freilich et al. where their resource competition is high enough contributing to their non-overlapping existence [20]. Moreover, similarly to commensalism, there is also the case when a taxon has a negative affect on the other without getting any harm (**amensalism**). Such is the case for *Acidithiobacillus thiooxidant* that produces sulfuric acid (H_2SO_4) by oxidation of sulfur [21] which is responsible for lowering of pH in the culture media which inhibits the growth of most other bacteria [22]. Finally, one of the taxa may have a positive affect (host) on the other, but the latter (parasite) can be harmful to its benefator (**parasitism**) [15]. There are multiple cases of parasitism in real-world communities; specis of the genus *Bdellovibrio* for example, are parasites of other (gram-negative) bacteria [23].

Apparently, the environmental conditions affect the ecological interactinos to a great extent. A pair of taxa may be competitors in one case but have a nutral intrection in another one. In addition, evolutionary processes may change certain interactions; for example moving from commensalism to parasitism [24]. Both ecological and environmental interactions play a part in the composition and the functional potential of microbial assemblages.

1.1.2 The omics' era

To discover the microbial taxa present in a sample, scientists have adopted multiple ways throught the years. It is but a particularly limited proportion of the microbial species can be cultured 2019. Therefore, monocultures and enrichment cultures allow us to observe only a small fraction of the actual diversity. As a consequence, other methods for the taxonomic identification of theses species was needed. development of different methods based on molecular analysis of microbial communities To address this challenge, scientists

1.1.3 Reverse ecology: transforming ecology into a high-throughput field

(from Roie Levy and Elhanan Borenstein [26]) Reverse Ecology—an emerging new frontier in Evolutionary Systems Biology—aims to extract this information and to obtain novel insights into an organism's ecology. The Reverse Ecology framework facilitates the translation of high-throughput genomic data into large-scale ecological data, and has the potential to transform ecology into a high-throughput field

(from RevEcoR publication [27]) A systematic approach for describing microbiome ecologies and the interactions between microbiota is lacking. To address this challenge, a systems biology approach called *reverse ecology* has been developed to study the complex interactions and species composition of microbial communities [4]. Reverse ecology uses genomics to study community ecology with no a priori assumptions about the organisms under consideration. Researchers can use it to infer the ecology of a system directly from genomic information. The reverse ecology framework uses advances in systems biology and genomic metabolic modeling and the system-level analysis of complex biological networks to predict the ecological traits of poorly studied microorganisms, their interactions with other microorganisms, and the ecology of microbial communities. Several studies have applied this approach to investigate the interactions between microorganisms and their surroundings on a large scale [4, 5].

The relationship between genotype and phenotype is fundamental to biology. Many levels of control are introduced when moving from one to the other. Systems biology aims at deciphering "the strategy" both at the cell and at higher levels of organization, in case of multicell species, that enables organisms to produce orderly adaptive behavior in the face of widely varying genetic and environmental conditions ([28]); the term "strategy" is used as per [29]. Systems biology approaches aim at interpreting how a system's properties emerge; from the cell to the community level. As Polanyi (1968) underlines "live mechanisms and information in DNA are boundary conditions with a sequence of boundaries above them".

1.2 Bioinformatics challenges in the analysis of HTS data

- need for tools
- handle the sequences

⟶ bridge to next section about data integration

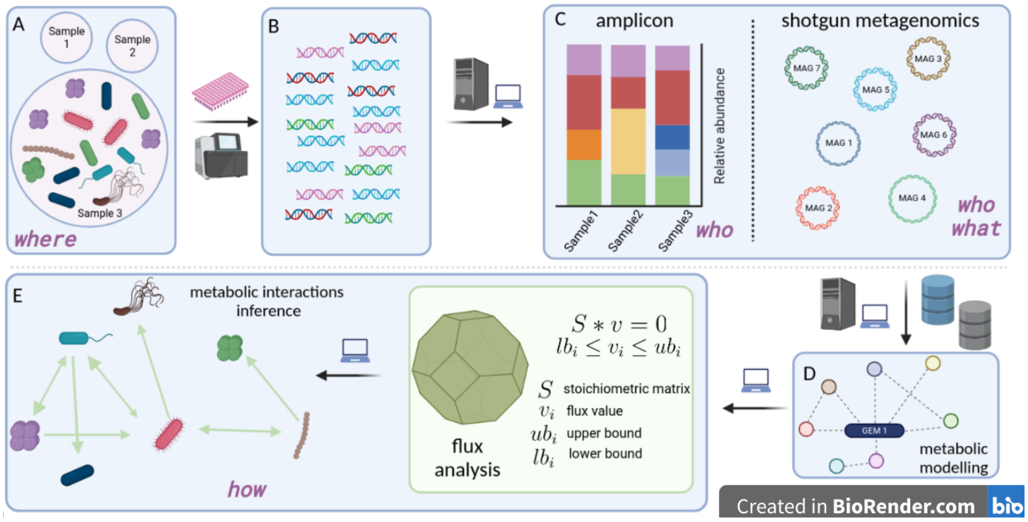

FIGURE 1.3:

1.3 Data integration & data mining in the service of microbial ecology

1.3.1 Metadata: a key issue for the microbiome community

The Community initially focused on developing open science "best practices" for the research community. The paper "The metagenomic data life-cycle: standards and best practices" [30] provided the foundation for FAIR data management in the domain. These best practices advocated using community standards for contextual provenance and metadata at all stages of the research data life cycle.

Alongside archived sequence data, access to comprehensive metadata is important to contextualise where the data originated. On submission, submitters are given the option to provide details regarding when, where and how their samples were collected with the opportunity to align provided metadata against community developed standards where possible. However, challenges associated with metadata deposition mean submitters do not always provide comprehensive metadata - these challenges can range from: lack of training and outreach resulting in submitters not fully understanding the importance of metadata and how to comply with standards; as well as the trade-offs for the archives to provide complex and thorough validation vs simple user interfaces to ensure both compliance and submission are as easy as possible. For the ENA, extensive documentation exists on how to submit data which both encourages compliance with metadata standards and provides separate submission guidelines for different data types - usage of the documentation can mitigate common errors and often aid first-time submitters but does not reach the full user-base.

FAIR principles, to provide a multilayer set of metadata required by the different scientific communities, reflecting the inherently multi-disciplinary character of environmental microbiology. The various layers of metadata necessary for the FAIRification of MAGs should include:

1. Environmental data describing the sample of origin
2. Sequencing technology or technologies
3. Details on the computational pipeline for metagenome assembly, binning and quality assessment
4. Connection to an existing taxonomy schema

OSD's open access strategy and provenance for metadata annotation is reflected in its ENA and Pangea submissions. Among others Standardization and training are key aspects across OSD: from sampling protocols to metadata checklists and guidelines. This is inline with aims of the Elixir microbiome community (see Sections "Mobilising raw data and metadata", "Training - lack of training"); spreading the experience to other biomes can benefit such ends.

Open questions: Metadata standard definition: minimum set and formats (Some flexibility will have to be considered in sharing standards between domain-specific communities). Systems to extract the vast amount of metadata locked in the scientific literature and provide them in standard format (explored by the Biodiversity Focus Group).

Metadata associated with the raw data, the assembled data, and the workflow. The necessary scripts will be written in Python using standard libraries and Biopython. Metadata of the cleaned data Metadata associated with the data sequencing, sample collection (MIMS), and quality control analyses will be generated according to the ENA manifest to enable uploading and archiving of the data to ENA. Metadata of the assembled data Because the workflow is distributed, it is necessary for EBI-MGnify to verify the provenance of the data workflow through registration and a verification test. A unique calculated hash generated from the data and workflow code will serve as a key for verification. This metadata will be generated at this step and together with the metadata associated with the assembly, uploaded to ENA/MGnify for further downstream functional annotation. Metadata to accompany the taxonomic inventories Metadata associated with the previous two steps will be summarised for inclusion with the taxonomic inventories (biom file format and CSV) for publication on the EMBRC GOs website.

- Metadata of the cleaned data; metadata associated with the data sequencing, sample collection (MIMS), and quality control analyses
- Metadata of the assembled data
- Metadata to accompany the taxonomic inventories

1.3.2 Ontologies & databases: the corner stone of mordern biology

Databases

- GenBank, ENA

- repositories such as MGnify
- PubMed

Ontologies:

- ENVO
- NCBI Taxonomy
- Gene Ontology
- Uniprot
- KEGG
- https://edamontology.org/page

1.4 Metabolic networks: modeling cellular physiology and growth

1.4.1 Genome-scale metabolic model analysis

Being at the helm of the most critical celular functions, metabolism and therefore, metabolic networks and their analysis, play a key role in Systems Biology. Moreover, Lewis et al. (2012) describe thoroughly the multiple constraint-based reconstruction and analysis (COBRA) methods that have been developed to support the analysis of such networks.

1.4.2 Sampling the flux space of a metabolic model: challenges & potential

1.5 *Reverse Ecology* and other Systems Biology - oriented methods to investigate life in extreme environments

1.6 Aims and objectives

The aim of this PhD was double:

1. to enhance the analysis of microbiome data by building algorithms and software to address some of the on-going computational challenges on the field.
2. to exploit these methods to identify taxa, functions, especially related to sulfur cycle, and microbial interactions that support life in microbial community assemblages in hypersaline sediments.

All parts of this work are purely computational. Samples and their corresponding sequencing data used in Chapter **??** have been collected and produced by Dr. Christina Pavloudi.

In **Chapter ??**, challenges derived from the analysis of HTS amplicon data are examined. A bioinformatics pipeline, called pema, for the analysis of several marker genes was developed, combinining several new technologies that allow large scale analysis of hundreds of samples. In addition, a software tool called darn, was built to investigate the unassigned sequences in amplicon data of the COI marker gene.

In **Chapter ??**, sediment samples from a hypersaline swamp in Tristomo, Karpathos Greece were analysed using both amplicon and shotgun metagenomics. The taxonomic and the functional profiles of the microbial communities present there were investigated. Key metabolic processes for ensuring life at such an extreme environment were identified. Microbial interactions of the assemblages retrieved were also studied.

In **Chapter ??**, data integration, data mining and text-mining methods were exploited to build a knowledge-base, called prego, including millions of associations between:

1. microbial taxa and the environments they have been found in
2. microbial taxa and biological processes they occur
3. environmental types and the biological processes that take place there

In **Chapter ??**, the challenges of flux sampling in metabolic models of high dimensions was presented along with a Multiphase Monte Carlo Sampling (MMCS) algorithm we developed.

In **Chapter ??**, the history of the IMBBC-HCMR HPC facility was presented indicating the vast needs of computing resources in modern analyses in general and in microbial studies more specifically.

Finally, in **Chapter ??**, general discussion and conclusions that have derived from this research were presented.

Appendices

Appendix A

Appendix: PREGO

A.1 Mappings

PREGO produces entity identifiers either by Named Entity Recognition (NER) with the EXTRACT tagger or by mapping retrieved identifiers to the selected ones. PREGO adopted NCBI taxonomy identifiers for taxa, Environmental Ontology for environments and Gene Ontology as a structure knowledge scheme for Processes (GObp) and Molecular Functions (GOmfs). The latter was for reasons that are two-fold, first Gene Ontology has a Creative Commons Attribution 4.0 License and second there are many resources that have mapped their identifiers to Gene Ontology. MG-RAST metagenomes and JGI/IMG isolates annotations come with KEGG orthology (KO) terms; Struo-oriented genome annotations, on the other hand, have Uniprot50 ids. The mapping from KO to GOmf and Uniprot50 to GOmf is implemented via UniProtKB mapping files of their FTP server (see idmapping.dat and idmapping_selected.tab files). By using the 3-column mapping file, the initial annotations were mapped to GOmf. As a complement, a list of metabolism-oriented KEGG ORTHOLOGY (KO) terms has been built (see *prego_mappings* in the Availability of Supporting Source Codes section). Finally, as STRUO annotations refer to GTDB genomes, publicly available mappings (accessed on 24 December 2021) were used to link the genomes used with their corresponding NCBI Taxonomy entries.

A.2 Daemons

An important component PREGO approach (Figure A1) is the regular updates which keep PREGO in line with the literature and microbiology data advances. The updates are implemented with custom scripts called daemons that are executed regularly spanning from once a month up to six-month cycles. This variation occurs because of the API requirements of each web resource as well as the computational intensity of the association extraction from the retrieved data.

Each Daemon is attached to a resource because its data retrieval methods (API, FTP) and following steps, shown in Figure A1, require special handling and multiple scripts (see *prego_daemons* in the Availability of Supporting Source Codes section).

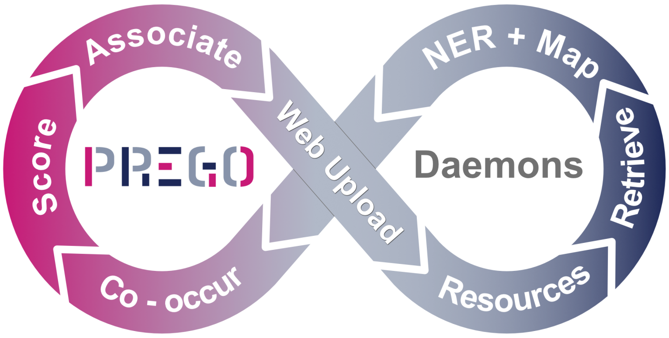

FIGURE A.1: Software daemons perform all steps of the PREGO methodology in a continuous manner similar to the Continuous Development and Continuous Integration method.

A.3 Scoring

Scoring in PREGO is used to answer the questions:

- Which associations are more thrustworthy?
- Which associations are more relevant to the user's query?

Relevant, informative, and probable associations are presented to the user through the three channels that were discussed previously. Each channel has its own scoring scheme for the associations it contains and all of them are fit in the interval $(0,5]$ to maintain consistency. The values of the score are visually shown as stars. The Genome Annotation and Isolates channel has fixed values of scores depending on the resource because Genome Annotation is straightforward, and the microbe id is known a priori. On the other hand, Environmental Samples channel data are based on samples, which contain metagenomes and OTU tables. Thus, it has two levels of organization, microbes with metadata, and sample identifiers. Each association of two entities is scored based on the number of samples they co-occur. A Literature channel scoring scheme is based on the co-mention of a pair of entities in each document, paragraph, and sentence. The differences in the nature of data require different scoring schemes in these channels. The contingency table (Table A.1) of two random variables, X and Y are the starting point for the calculation of scores. The term $X = 1$ might be a specific NCBI id and $Y = 1$ a ENVO term. The $c_{1,1}$ is the number of instances that two terms of $X = 1$ and $Y = 1$ are co-occurring, i.e., the joint frequency. The marginals are the $c_{1,.}$ and $c.,1$ for x and y, respectively, which are the backgrounds for each entity type. Different handling of these frequencies leads to different measures. There is not a perfect scoring scheme, just the one that works best on a particular instance. Consequently, scoring attributes require testing different measures and their parameters.

		Y = y		
		Yes	No	Total
X = x	Yes	$c_{x,y}$	$c_{x,0}$	$c_{x,.}$
	No	$c_{0,y}$	$c_{0,0}$	$c_{0,.}$
	Total	$c_{.,y}$	$c_{.,0}$	$c_{.,.}$

TABLE A.1: Contingency table of co-occurrences between entities $X = x$ and $Y = y$. This is the basic structure for all scoring schemes. $c_{x,y}$ is the count of the co-occurrence of these entities. $c_{x,.}$ is the count of the x with all the entities of Y type (e.g., Molecular function). Conversely, $c_{.,y}$ is the count of y with all the entities of X type (e.g., taxonomy

Literature Channel

Scoring in the Literature channel is implemented as in STRING 9.1 [32] and COMPARTMENTS [33], where the text mining method uses a three-step scoring scheme. First, for each co-mention/co-occurrence between entities (e.g., Methanosarcina mazei with Sulfur carrier activity), a weighted count is calculated because of the complexity of the text.

$$c_{x,y} = \sum_{k=1}^{n} w_d \delta_{dk}(x,y) + w_p \delta_{p,k}(x,y) + w_s \delta_{sk}(x,y) \tag{A.1}$$

Different weights are used for each part of the document (k) for which both entities have been co-mentioned, $w_d = 1$ for the weight for the whole document level, $w_p = 2$ for the weight of the paragraph level, and $w_s = 0.2$ for the same sentence weight. Additionally, the delta functions are one (Equation A.1) in cases the co-mention exists, zero otherwise. Thus, the weighted count becomes higher as the entities are mentioned in the same paragraph and even higher when in the same sentence. Subsequently, the co-occurrence score is calculated as follows:

$$score_{x,y} = c^{a}_{x,y} \left(\frac{c_{x,y} c_{.,.}}{c_{x,.} c_{.,y}}\right)^{1-a} \tag{A.2}$$

where $a = 0.6$ is a weighting factor, and the $c_{x,.}$, $c_{.,.}$, $c_{.,y}$ are the weighted counts as shown in Table A.1 estimated using the same Equation A.2. This value of the weighting factor has been chosen because it has been optimized and benchmarked in various applications of text mining [34,70,71]. The value of Equation A.2 is sensitive to the increasing size of the number of documents (MEDLINE PubMed—PMC OA). Therefore, to obtain a more robust measure, the value of the score is transformed to z-score. This transformation is elaborated in detail in the COMPARTMENTS resource [33]. Finally, the confidence score is the z-score divided by two. Cases in which the scores exceed the (0,4] interval are capped to a maximum of 4 to reflect the uncertainty of the text mining pipeline.

Environmental Samples Channel

Data from environmental samples are OTU tables and metagenomes. Thus, for each entity x, the number of samples is calculated as the background and a number of samples of the associated entity (metadata background) c.,y (see Table A1). Each association between entities x, y has a number of samples, cx,y that they co-occur. Note that each resource is independent and the scoring scheme is applied to its entities. This means that the same association can appear in multiple resources with different scores. The score is calculated with the following formula:

$$score_{x,y} = 2.0 * \sqrt{\frac{c_{x,y}}{c_{.,y}}}^{a} \tag{A.3}$$

This score is asymmetric because the denominator is the marginal of the associated entity. Thus, the score decreases as the marginal of y is increasing, i.e., the number of samples that y is found. On the other hand, it promotes associations in which the number of samples of the association are similar to the marginal of y. The exponents on the numerator and denominator equal to 0.5 and to 0.1, respectively, in order to reduce the rapid increase of score. Lastly, the value of the score is capped in the range $(0,4]$.

A.4 Bulk download

Users can also download programmatically all associations per channel through the links that are shown in Table A.2. The data are compressed to reduce the download size and md5sum files are provided as well for a sanity check of each download.

Channel	Link	md5sum	Size (in GB)
Literature	literature.tar.gz	literature.tar.gz.md5	5.4
Environmental Samples	environmental_samples.tar.gz	environmental_samples.tar.gz.md5	0.69
Annotated genomes and isolates	annotated_genomes_isolates.tar.gz	annotated_genomes_isolates.tar.gz.md5	0.26

TABLE A.2: Bulk download links and md5sum files.

Bibliography

[1] Paul G Falkowski, Tom Fenchel, and Edward F Delong. The microbial engines that drive earth's biogeochemical cycles. *science*, 320(5879):1034–1039, 2008.

[2] Jodie Belilla, David Moreira, Ludwig Jardillier, Guillaume Reboul, Karim Benzerara, José M López-García, Paola Bertolino, Ana I López-Archilla, and Purificación López-García. Hyperdiverse archaea near life limits at the polyextreme geothermal dallol area. *Nature ecology & evolution*, 3(11):1552–1561, 2019.

[3] Kenneth J Locey and Jay T Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, 2016.

[4] What is microbial ecology? URL https://www.isme-microbes.org/what-microbial-ecology.

[5] Paul V Dunlap. Microbial diversity. 2001.

[6] MT Madigan, KS Bender, DH Buckley, WM Sattley, and DA Stahl. Brock biology of microorganisms. 15th global edition. *Boston, US: Benjamin Cummins*, 2018.

[7] Stilianos Louca, Laura Wegener Parfrey, and Michael Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016.

[8] Gabriel E Leventhal, Carles Boix, Urs Kuechler, Tim N Enke, Elzbieta Sliwerska, Christof Holliger, and Otto X Cordero. Strain-level diversity drives alternative community types in millimetre-scale granular biofilms. *Nature microbiology*, 3(11):1295–1303, 2018.

[9] Maria L Marco. Defining how microorganisms benefit human health. *Microbial Biotechnology*, 14(1):35–40, 2021.

[10] Ricardo Cavicchioli, William J Ripple, Kenneth N Timmis, Farooq Azam, Lars R Bakken, Matthew Baylis, Michael J Behrenfeld, Antje Boetius, Philip W Boyd, Aimée T Classen, et al. Scientists' warning to humanity: microorganisms and climate change. *Nature Reviews Microbiology*, 17(9):569–586, 2019.

[11] Samir Giri, Leonardo Oña, Silvio Waschina, Shraddha Shitut, Ghada Yousif, Christoph Kaleta, and Christian Kost. Metabolic dissimilarity determines the establishment of cross-feeding interactions in bacteria. *Current Biology*, 31(24):5547–5557, 2021.

[12] Samir Giri, Shraddha Shitut, and Christian Kost. Harnessing ecological and evolutionary principles to guide the design of microbial production consortia. *Current Opinion in Biotechnology*, 62:228–238, 2020.

[13] Wentao Kong, David R Meldgin, James J Collins, and Ting Lu. Designing microbial consortia with defined social interactions. *Nature Chemical Biology*, 14(8):821–829, 2018.

[14] Raíssa Mesquita Braga, Manuella Nóbrega Dourado, and Welington Luiz Araújo. Microbial interactions: ecology in a molecular perspective. *Brazilian Journal of Microbiology*, 47:86–98, 2016.

[15] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012.

[16] Alan R Pacheco, Mauricio Moel, and Daniel Segrè. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature communications*, 10(1):1–12, 2019.

[17] Alfonso Santos-Lopez, Christopher W Marshall, Michelle R Scribner, Daniel J Snyder, and Vaughn S Cooper. Evolutionary pathways to antibiotic resistance are dependent upon environmental structure and bacterial lifestyle. *Elife*, 8:e47612, 2019.

[18] Hendrikus J Laanbroek, Marie-Josee Baar-Gilissen, and Hans L Hoogveld. Nitrite as a stimulus for ammonia-starved nitrosomonas europaea. *Applied and Environmental Microbiology*, 68(3):1454–1457, 2002.

[19] Octavio Perez-Garcia, Gavin Lear, and Naresh Singhal. Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Frontiers in microbiology*, 7:673, 2016.

[20] Shiri Freilich, Anat Kreimer, Isacc Meilijson, Uri Gophna, Roded Sharan, and Eytan Ruppin. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic acids research*, 38(12):3857–3868, 2010.

[21] Roberto A Bobadilla Fazzini, Maria Paz Cortés, Leandro Padilla, Daniel Maturana, Marko Budinich, Alejandro Maass, and Pilar Parada. Stoichiometric modeling of oxidation of reduced inorganic sulfur compounds (riscs) in acidithiobacillus thiooxidans. *Biotechnology and Bioengineering*, 110(8):2242–2251, 2013.

[22] Qusheng Jin and Matthew F Kirk. ph as a primary control in environmental microbiology: 1. thermodynamic perspective. *Frontiers in Environmental Science*, 6:21, 2018.

[23] H Stolp. Interactions between bdellovibrio and its host cell. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 204(1155):211–217, 1979.

[24] Eric Parmentier, Déborah Lanterbecq, and Igor Eeckhaut. From commensalism to parasitism in carapidae (ophidiiformes): heterochronic modes of development? *PeerJ*, 4:e1786, 2016.

[25] Andrew D Steen, Alexander Crits-Christoph, Paul Carini, Kristen M DeAngelis, Noah Fierer, Karen G Lloyd, and J Cameron Thrash. High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal*, 13(12):3126–3130, 2019.

[26] Roie Levy and Elhanan Borenstein. Reverse ecology: from systems to environments and back. In *Evolutionary systems biology*, pages 329–345. Springer, 2012.

[27] Yang Cao, Yuanyuan Wang, Xiaofei Zheng, Fei Li, and Xiaochen Bo. Revecor: an r package for the reverse ecology analysis of microbiomes. *BMC bioinformatics*, 17(1):1–6, 2016.

[28] Richard Strohman. Maneuvering in the complex path from genotype to phenotype. *Science*, 296(5568):701–703, 2002.

[29] Michael Polanyi. Life's irreducible structure: Live mechanisms and information in dna are boundary conditions with a sequence of boundaries above them. *Science*, 160(3834):1308–1312, 1968.

[30] Petra Ten Hoopen, Robert D Finn, Lars Ailo Bongo, Erwan Corre, Bruno Fosso, Folker Meyer, Alex Mitchell, Eric Pelletier, Graziano Pesole, Monica Santamaria, et al. The metagenomic data life-cycle: standards and best practices. *GigaScience*, 6(8):gix047, 2017.

[31] Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291–305, 2012.

[32] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2012.

[33] Janos X Binder, Sune Pletscher-Frankild, Kalliopi Tsafou, Christian Stolte, Seán I O'Donoghue, Reinhard Schneider, and Lars Juhl Jensen. Compartments: unification and visualization of protein subcellular localization evidence. *Database*, 2014, 2014.

Short CV

Education

- **Doctor of Philosophy** (2018 – 2022), University of Crete, Biology Department
  Thesis: Microbial communities through the lens of high throughput sequencing, data integration and metabolic networks analysis
  Thesis conducted at IMBBC - HCMR

- **M.Sc. in Bioinformatics** (2016 – 2018), University of Crete, School of Medicine
  Thesis: eDNA metabarcoding for biodiversity assessment: Algorithm design and bioinformatics analysis pipeline implementation
  Thesis conducted at IMBBC - HCMR

- **B.Sc. in Biology** (2011 – 2016), National and Kapodistrian University of Athens, department of Biology
  Thesis: Morphology, morphometry and anatomy of species of the genus *Pseudamnicola* in Greece

Research projects - working Experience

- **A workflow for marine Genomic Observatories data analysis** (2021 - ongoing)
  Role: scientific responsiblse & developer
  This EOSC-Life funded project aims at developing a workflow for the analysis of EMBRC's Genomic Observatories (GOs) data, allowing researchers to deal better with this increasing amount of the data and make them more easily interpretable.

- **PREGO: Process, environment, organism (PREGO)** (2019 - 2021)
  Role: PhD candidate
  PREGO is a systems-biology approach to elucidate ecosystem function at the microbial dimension.

- **ELIXIR-GR** (2019 - 2021)
  Role: technical support
  ELIXIR-GR is the Greek National Node of the ESFRI European RI ELIXIR, a distributed e-Infrastructure aiming at the construction of a sustainable European infrastructure for biological information.

- **RECONNECT** (2018 - 2020)
  Role: technical support
  RECONNECT is an Interreg V-B "Balkan-Mediterranean 2014-2020" project. It aims to develop strategies for sustainable management of Marine Protected Areas (MPAs) and Natura 2000 sites.

Publications

- **PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types.**
  Zafeiropoulos, H., Paragkamian S.[2], Stelios Ninidakis, Georgios A. Pavlopoulos, Lars Juhl Jensen, and Evangelos Pafilis. *Microorganisms* 10, no. 2 (2022): 293., DOI: 10.3390/microorganisms10020293

- **The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data**
  Zafeiropoulos H., Gargan L., Hintikka S., Pavloudi C., & Carlsson J. *Metabarcoding and Metagenomics*, 5, p.e69657, 2021, DOI: 10.3897/mbmg.5.69657

- **0s & 1s in marine molecular research: a regional HPC perspective.**
  Zafeiropoulos H., Gioti A., Ninidakis S., Potirakis A., ..., & Pafilis E. *GigaScience*, 9(3), p.giab053, 2021 DOI: 10.1093/gigascience/giab053

- **Geometric Algorithms for Sampling the Flux Space of Metabolic Networks**
  Chalkis, A., Fisikopoulos, V., Tsigaridas, E. & **Zafeiropoulos, H.** *37th International Symposium on Computational Geometry (SoCG 2021)*, 21:1–21:16, 189, 2021 DOI: 10.4230/LIPIcs.SoCG.2021.21

- **The Santorini Volcanic Complex as a Valuable Source of Enzymes for Bioenergy**
  Polymenakou, P.N., Nomikou, P., **Zafeiropoulos, H.**, Mandalakis, M., Anastasiou, T.I., Kilias, S., Kyrpides, N.C., Kotoulas, G. & Magoulas,A. *Energies*, 14(5), p.1414, 2021 DOI: 10.3390/en14051414

- **PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes**
  Zafeiropoulos, H., Viet, H.Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C. & Pafilis, E. *GigaScience*, 9(3), p.giaa022, 2020 DOI: 10.1093/gigascience/giaa022

In preparation

- dingo: a Python library for metabolic networks analysis

- Deciphering the functional potential of a hypersaline swamp microbial mat community

[2] ZH and PS contibuted equally in this study

Awards

- **European Molecular Biology Organization Short-Term Fellowship** (2022)
  Project title: Exploiting data integration, text-mining and computational geometry to enhance microbial interactions inference from co-occurrence networks
  https://hariszaf.github.io/microbetag/

- **Mikrobiokosmos travel grant in memorium of Prof. Kostas Drainas** (2021)

- **Google Summer of Code** (2021)
  Project title: From DNA sequences to metabolic interactions: building a pipeline to extract key metabolic processes
  Report, GSoC archive

- **Federation of European Microbiological Societies Meeting Attendance Grant** (2020)
  for joining the *Metagenomics, Metatranscript- omics and multi 'omics for microbial community studies* Physalia course

- **Short Term Scientific Mission (STSM) - DNAqua-net COST action** (2019)
  Project title: A comparison of bioinformatic pipelines and sampling techniques to enable benchmarking of DNA metabarcoding
  Report

- **Best Poster Award @ Hellenic Bioinformatics conference** (2018)
  for *PEMA: a Pipeline for Environmental DNA Metabarcoding Analysis*

Selected presentations

- **Bioinformatics Open Source Conference - BOSC2021** (2021)
  dingo: A python library for metabolic networks sampling & analysis, video poster - video

- **1st DNAQUA International Conference** (2021)
  PEMA v2: addressing metabarcoding bioinformatics analysis challenges, oral talk - video

- **Federation of European Microbiological Societies - FEMS2020** (2020)
  "Mining literature and -omics (meta)data to associate microorganisms, biological processes and environment types" - video poster

- **PyData Global PyData2020**
  "Geometric and statistical methods in systems biology: the case of metabolic networks", oral talk - video

- **8th International Barcode of Life Conference** - 2019
  "P.E.M.A.: a pipeline for environmental DNA metabarcoding analysis" (flashtalk)

Participation in proposal writing

- "Climate Change Metagenomic Record Index (CCMRI)" project: submitted at the 2nd Call for H.F.R.I Research Projects to Support Faculty Members & Researchers (June 2020). Approved for funding
- "A workflow for marine Genomic Observatories data analysis" project: submitted at the second Training Open Call of EOSC-Life (November 2020). Approved for funding

Contact

You may find more about me and my work on my personal website.
You can also find me on GitHub, Twitter and ResearchGate.
Email: haris-zaf@hcmr.gr, haris.zafr@gmail.com