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. Panagiotis Sarris
University of Crete
Department of Biology

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

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

Haris Zafeiropoulos

Contents

Preface i

Contents ii

Abstract v

Περίληψη vii

List of Figures and Tables viii

List of Abbreviations and Symbols x

1 Introduction 1

1.1 Microbial ecology . . . 1

1.1.1 Microbial communities: structure & function . . . 1

1.1.2 The role of microbial communities in biogeochemical cycles . . . 1

1.1.3 Microbial interactions: unravelling the microbiome . . . 1

1.2 The era of omics . . . 2

1.2.1 High Throughput Sequencing approaches . . . 2

1.2.2 Bioinformatics challenges . . . 2

1.3 Data integration & data mining in the era of omics . . . 3

1.3.1 Metadata: a key issue for the microbiome community . . . 3

1.3.2 Ontologies & databases: the corner stone of mordern biology . . . 5

1.4 Metabolic modeling at the omics era . . . 6

1.4.1 Genome-scale metabolic model analysis . . . 6

1.4.2 Sampling the flux space of a metabolic model: challenges & potential 6

1.5 The hypersaline Tristomo swamp: a case study of an extreme environment 6

1.6 Systems biology from a computational resources point-of-view . . . 6

1.7 Aims and objectives . . . 6

2 Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment 8

2.1 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes . . . 8

2.1.1 Abstract . . . 8

2.1.2 Introduction . . . 9

2.1.3 Contribution . . . 10

2.1.4 Methods & Implementation . . . 11

2.1.5 Results & Validation . . . 14

2.1.6 Discussion . . . 21
2.1.7 Supplementary Material . . . 23
2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data . . . 25
2.2.1 Abstract . . . 25
2.2.2 Introduction . . . 25
2.2.3 Contribution . . . 27
2.2.4 Methods & Implementation . . . 27
2.2.5 Results & Validation . . . 31
2.2.6 Discussion . . . 35

3 Studying the microbiome as a whole: the way forward 36
3.1 Microbial interactions inference in communities of a hypersaline swamp elucidate mechanisms governing taxonomic & functional profiles . . . 36
3.1.1 Introduction . . . 36
3.1.2 Contribution . . . 36
3.1.3 Methods . . . 36
3.1.4 Results . . . 36
3.1.5 Discussion . . . 36

4 Software development to build a knowledge-base at the systems biology level 37
4.1 PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types . . . 37
4.1.1 Abstract . . . 37
4.1.2 Introduction . . . 38
4.1.3 Contribution . . . 40
4.1.4 Methods & Implementation . . . 40
4.1.5 Results & Validation . . . 44
4.1.6 Discussion . . . 50

5 Software development to establish metabolic flux sampling approaches at the community level 56
5.1 A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks 56
5.1.1 Abstract . . . 56
5.1.2 Introduction . . . 57
5.1.3 Contribution . . . 62
5.1.4 Methods & Implementation . . . 63
5.1.5 Results . . . 69
5.1.6 Parameter tuning for practical performance . . . 69
5.1.7 Experiments . . . 72
5.2 Conclusions and future work . . . 74

6 An overview of the computational requirements & solutions in microbial ecology 75
6.1 0s and 1s in marine molecular research: a regional HPC perspective . . . 75
6.1.1 Abstract . . . 75
6.1.2 Introduction . . . 76

6.1.3 Contribution 77
6.1.4 Methods 78
6.1.5 Results 81
6.1.6 Discussion 84
6.1.7 Conclusions 91

7 Conclusions **92**

A Appendix: PREGO **95**
A.1 Mappings 95
A.2 Daemons 95
A.3 Scoring 96
A.4 Bulk download 98

Bibliography **99**

Short CV **124**

Abstract

Microbial communities are a cornerstone for most ecosystem types. To elucidate the mecahinsms governing such assemblages, it is fundamental to identify the taxa present (*who*) and the processeses 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. But also, to exploit state-of-the-art methods to in identify taxa, functions and microbial interactions in assemblages of various aquatic environmnets. To this end, a number of publically available datasets 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 infustructures ease this need to a great extent and PEMA was developed to this end (Chapter 2.1). Moreover, 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 OTUs/ASVs returned 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 2.2).

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 analysed. 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 3).

Both 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 4).

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 5). GEMs were built using the MAGs retireved 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

List of Figures and Tables

List of Figures

1.1 The cycle of C and the role of microbial communiites . . . 2
1.2 Sulfur cycle . . . 3
1.3 Nitrogen cycle . . . 4

2.1 PEMA in a nutshell . . . 11
2.2 Phylogeny - based taxonomy assignment in PEMA . . . 13
2.3 OTU bar plot at the phylum level. . . . 20
2.4 Building the COI reference tree of life . . . 29
2.5 Placements of the consensus COI sequences on the reference COI tree . . . 32

4.1 PREGO analysis methodology . . . 40
4.2 PREGO web user interface . . . 45
4.3 PREGO in action - examples . . . 46
4.4 The PREGO API schema . . . 47
4.5 Summary of the unique entities per phylum for each of the four entity types on PREGO . . . 51

5.1 From DNA sequences to distributions of metabolic fluxes . . . 59
5.2 Flux distributions in the most recent human metabolic network Recon3D . . . 60
5.3 A Multiphase Monte Carlo Sampling algorithm . . . 67
5.4 Estimation, using our tools, of the marginal distribution of 6 reaction fluxes in two constraint-based model of Homo Sapiens metabolism, namely Recon2D [1] (blue color) and Recon3D [2] (red color). In the case of GK1 we observe how the flux distribution of a reaction may change once the direction of the reaction changes. . . . 70

6.1 The IMBBC HPC facility history . . . 78
6.2 Block diagram of the *Zorba* architecture. . . . 79
6.3 IMBBC HPC supported published studies grouped by scientific field . . . 82
6.4 Computational resources requirements of the so-far published studies supported by the IMBBC HPC facility . . . 83

A.1 PREGO DevOps . . . 96

List of Tables

2.1 Summary benchmark of PEMA marker - gene - specific mock community recovery . . . 16

2.2 Comparison of the basic features of the different metabarcoding bioinformatics pipelines . . . 18

2.3 OTU predictions and execution time for the different pipelines . . . 19

2.4 PEMA's output and execution time . . . 21

2.5 Comparing taxonomies retrieved from PEMA and Barque pipelines . . . 22

2.6 Number of sequences and taxonomic species per domain of life and resources 28

2.7 DARN outcome over the samples or set of samples . . . 34

4.1 Source databases integrated in PREGO and the number of items retrieved . . . 41

4.2 The entities of PREGO after the NER and mapping of every source . . . 49

4.3 Associations among the PREGO entities . . . 52

4.4 Feature comparison between PREGO and other similar platforms . . . 53

5.1 Recon2 and Recond3D distribution comparison . . . 71

5.2 MMCS time and PSRF per phase . . . 73

5.3 Sampling from iAF1260 . . . 74

A.1 PREGO contingency table between two terms . . . 97

A.2 PREGO Bulk download links and md5sum files. . . . 98

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

Chapter 1

Introduction

1.1 Microbial ecology

1.1.1 Microbial communities: structure & function

It was only 2 years ago that scientists managed to find a place on Earth where no microbial forms of life are present; this was possible thanks to to very low pH, high salt and high temperature a [3]

Microbes, i.e. Bacteria, Archaea and small Eukaryotes such as protoza, are omnipresent and impact global ecosystem functions [4] through their abundance [5], versatility [6] and interactions [7].

1.1.2 The role of microbial communities in biogeochemical cycles

Microbial communities at hydrothermal vents mediate the transformation of energy and minerals produced by geological activity into organic material. Organic matter produced by autotrophic bacteria is then used to support the upper trophic levels. The hydrothermal vent fluid and the surrounding ocean water is rich in elements such as iron, manganese and various species of sulfur including sulfide, sulfite, sulfate, elemental sulfur from which they can derive energy or nutrients.[8] Microbes derive energy by oxidizing or reducing elements. Different microbial species use different chemical species of an element in their metabolic processes. For example, some microbe species oxidize sulfide to sulfate and another species will reduce sulfate to elemental sulfur. As a result, a web of chemical pathways mediated by different microbial species transform elements such as carbon, sulfur, nitrogen, and hydrogen, from one species to another. Their activity alters the original chemical composition produced by geological activity of the hydrothermal vent environment.[9]

1.1.3 Microbial interactions: unravelling the microbiome

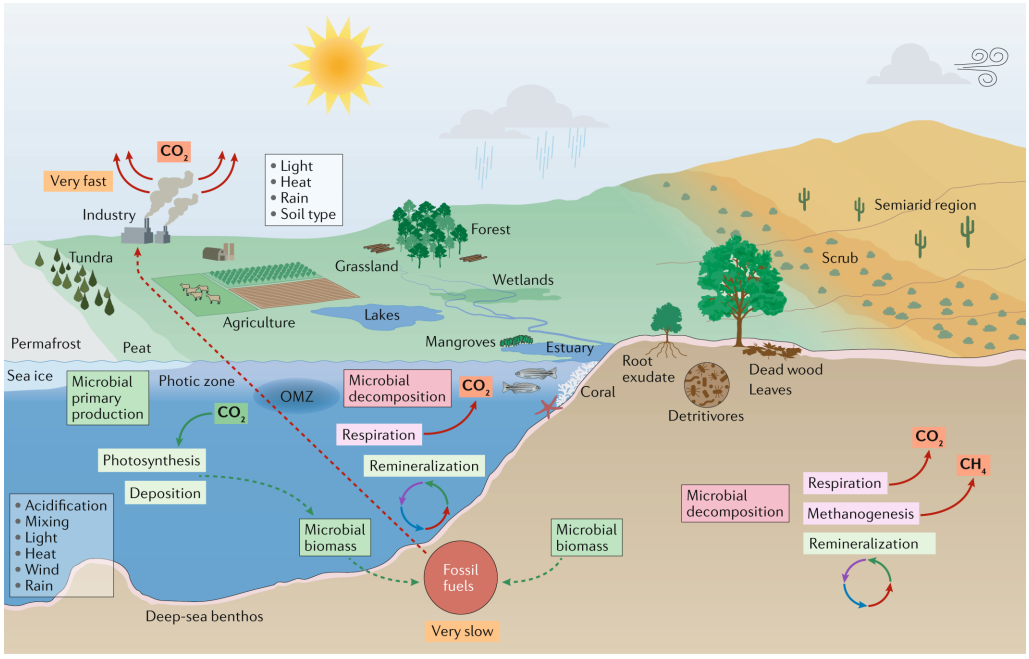

FIGURE 1.1: Marine microbial communities contribute to CO_2 sequestration, nutrients recycle and thus to the release of CO_2 to the atmosphere. Soil microbial communities decomposers organic matter and release nutrients in the soil from [8] doi: 10.1038/s41579-019-0222-5, under Creative Commons Attribution 4.0 International License

1.2 The era of omics

1.2.1 High Throughput Sequencing approaches

DNA metabarcoding is a rapidly evolving method that is being more frequently employed in a range of fields, such as biodiversity, biomonitoring, molecular ecology and others [11, 12]. Environmental DNA (eDNA) metabarcoding, targeting DNA directly isolated from environmental samples (e.g., water, soil or sediment, [13]), is considered a holistic approach [14] in terms of biodiversity assessment, providing high detection capacity. At the same time, it allows wide scale rapid bio-assessment [14] at a relatively low cost as compared to traditional biodiversity survey methods [15]. The underlying idea of the method is to take advantage of genetic markers, i.e. marker loci, using primers anchored in conserved regions. These universal markers should have enough sequence variability to allow distinction among related taxa and be flanked by conserved regions allowing for universal or semi-universal primer design [16].

1.2.2 Bioinformatics challenges

- need for tools
- handle the sequences

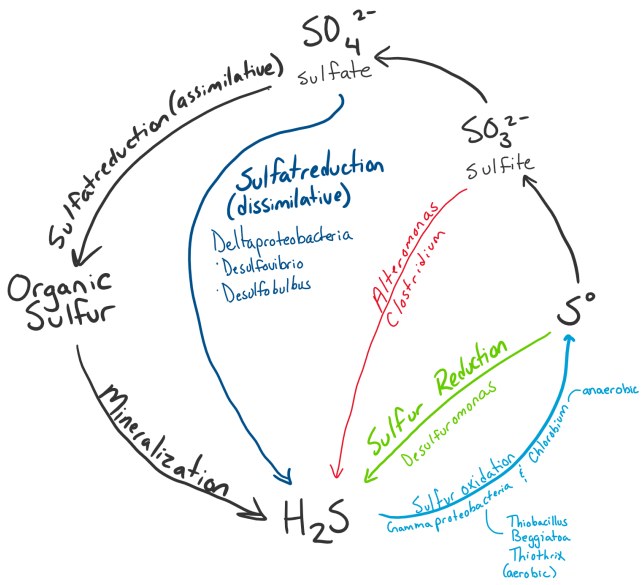

FIGURE 1.2: Sulfur cycle. Figure taken from [9]

1.3 Data integration & data mining in the era of omics

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" [17] 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 scien-

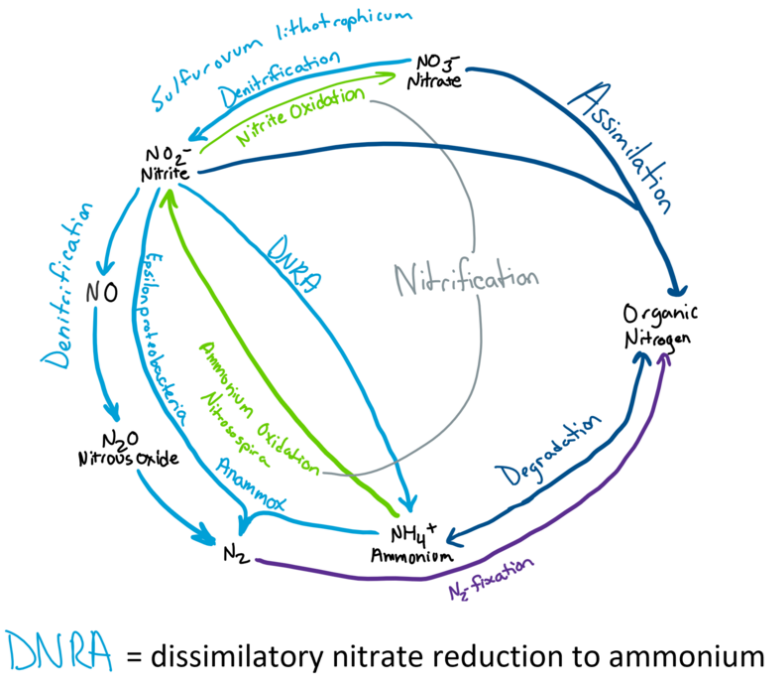

FIGURE 1.3: Nitrogen cycle. Figure taken from [10]

tific 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 modeling at the omics era

1.4.1 Genome-scale metabolic model analysis

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 ([18]); the term "strategy" is used as per [19]. Systems biology approaches aim at interpreting how a system's properties emerge; from the cell to the community level.

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

1.5 The hypersaline Tristomo swamp: a case study of an extreme environment

1.6 Systems biology from a computational resources point-of-view

1.7 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 3 have been collected and produced by Dr. Christina Pavloudi.

In **Chapter 2**, 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 3**, 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 4**, 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 5**, 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 microbial interactions for the assemblages were infered. All the methods developed and presented in the previous chapters were used to enhance the analysis of this microbiome.

In **Chapter 6**, 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 the **Conclusions** chapter, general discussion and conclusions that have derived from this research were presented.

Chapter 2

Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment

2.1 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes[1]

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

2.1.1 Abstract

Background: Environmental DNA and metabarcoding allow the identification of a mixture of species and launch a new era in bio- and eco-assessment. Many steps are required to obtain taxonomically assigned matrices from raw data. For most of these, a plethora of tools are available; each tool's execution parameters need to be tailored to reflect each experiment's idiosyncrasy. Adding to this complexity, the computation capacity of high-performance computing systems is frequently required for such analyses. To address the difficulties, bioinformatic pipelines need to combine state-of-the art technologies and algorithms with an easy to get-set-use framework, allowing researchers to tune each study. Software containerization technologies ease the sharing and running of software packages

[1] For author contributions, please refer to the relevant section. Modified version of the published review; extra features have been added and discussed on this thesis.

across operating systems; thus, they strongly facilitate pipeline development and usage. Likewise programming languages specialized for big data pipelines incorporate features like roll-back checkpoints and on-demand partial pipeline execution.

Findings: PEMA is a containerized assembly of key metabarcoding analysis tools that requires low effort in setting up, running, and customizing to researchers' needs. Based on third-party tools, PEMA performs read pre-processing, (molecular) operational taxonomic unit clustering, amplicon sequence variant inference, and taxonomy assignment for 16S and 18S ribosomal RNA, as well as ITS and COI marker gene data. Owing to its simplified parameterization and checkpoint support, PEMA allows users to explore alternative algorithms for specific steps of the pipeline without the need of a complete re-execution. PEMA was evaluated against both mock communities and previously published datasets and achieved results of comparable quality.

Conclusions: A high-performance computing–based approach was used to develop PEMA; however, it can be used in personal computers as well. PEMA's time-efficient performance and good results will allow it to be used for accurate environmental DNA metabarcoding analysis, thus enhancing the applicability of next-generation biodiversity assessment studies.

2.1.2 Introduction

Environmental DNA (eDNA) metabarcoding inaugurates a new era in bio- and eco-monitoring [20]. eDNA refers to genetic material obtained directly from environmental samples (soil, sediment, water, etc.) without any obvious signs of biological source material [21]. Metabarcoding is the combination of DNA taxonomy, based on taxa-specific marker genes (e.g., 16S ribosomal RNA [rRNA] for Bacteria and Archaea, cytochrome oxidase subunit 1 [COI] and 18S rRNA for Metazoa, ITS for Fungi), and high-throughput DNA sequencing technologies; thus, simultaneous identification of a mixture of organisms is attainable [15]. eDNA metabarcoding attempts to turn the page on the way biodiversity is perceived and monitored [15]. This combination is considered to be a potential holistic approach that, once standardized, allows for higher detection capacity and at a lower cost compared to conventional methods of biodiversity assessment. However, from the raw read sequence files to an amplicon study's results, the bioinformatics analysis required can be troublesome for many researchers.

Well-established pipelines are available to process metabarcoding data for the case of 16S and 18S rRNA marker genes and bacterial communities (e.g., mothur [22], QIIME 2 [23], LotuS [24]). However, certain limitations accompany each of these and occasionally they can be far from easy-to-use software. Moreover, there is a great need for similarly straightforward and benchmarked approaches for the analysis of other marker genes. With respect to the COI and ITS marker genes, a number of pipelines have been implemented, e.g., Barque[2], ScreenForBio [25], and PIPITS [26]. However, there is still need for a fast, flexible, easy-to-install, and easy-to-use pipeline for both COI and ITS marker genes.

The pipelines mentioned above, although entrenched, are still hindered by a series of hurdles. Among the most prominent are technical difficulties in installation and use,

[2] https://github.com/enormandeau/barque

strict limitations in setting parameters for the algorithms invoked, and incompetence in partial re-execution of an analysis.

Moreover, given the computational demands of such analyses, access to high - performance computing (HPC) systems might be mandatory, e.g., to process studies with a large number of samples. This is timely given the ongoing investment of national and international efforts (e.g., see European Strategy Forum on Research Infrastructures[3]) to serve the broad biological community via commonly accessible infrastructures.

2.1.3 Contribution

PEMA (Pipeline for Environmental DNA Metabarcoding Analysis) is an open source pipeline that bundles state-of-the-art bioinformatic tools for all necessary steps of amplicon analysis and aims to address the aforementioned issues. It is designed for paired-end sequencing studies and is implemented in the BDS [27] programming language. BDS's ad hoc task parallelism and task synchronization supports heavyweight computation, which PEMA inherits. In addition, BDS supports "checkpoint" files that can be used for partial re-execution and crash recovery of the pipeline. PEMA builds on this feature to serve tool and parameter exploratory customization for optimal metabarcoding analysis fine tuning. Switching effortlessly between (molecular) operational taxonomic unit ([M]OTU) clustering and amplicon sequence variant (ASV) inference algorithms is a pertinent example. Finally, via software containerization technologies such as Docker [28] and Singularity [29], with the latter being HPC-centered, PEMA is distributed in an easy to download and install fashion on a range of systems, from regular computers to cloud or HPC environments.

From the biological perspective, monitoring biodiversity at all its different levels is of great importance. Because there is not a single marker gene to detect all taxa, researchers need to use different genes targeting each great taxonomy group separately [30]. To that end, PEMA supports the metabarcoding analysis of both prokaryotic communities, based on the 16S rRNA marker gene, and eukaryotic ones, based on the ITS (for Fungi) and COI and 18S rRNA (for Metazoa) marker genes [30].

As high-throughput sequencing (HTS) data become more and more accurate, ASVs, i.e., marker gene amplified sequence reads that differ in ≥ 1 nucleotide from each other, become easier to resolve [31]. The use of ASVs instead of OTUs has been suggested [31]; however, the choice of which approach to use should be based on each study's objective(s) [32].

PEMA supports both OTU clustering and ASV inference for all marker genes (see "OTU clustering vs ASV inference" in the "Results and Discussion" section). Two clustering algorithms, VSEARCH [33] and CROP [34], are used for the clustering of reads in (M)OTUs—the former for the case of the 16S/18S rRNA marker genes, the latter for the case of COI and ITS. Swarm v2 [35] allows ASV inference in all cases.

Taxonomic assignment is performed in an alignment-based approach, making use of the CREST LCAClassifier [36] and the Silva database [37] for the case of 16S and 18S rRNA marker genes; the Unite database [38] is used for the ITS gene. In the 16S marker gene case,

[3] https://www.esfri.eu/sites/default/files/u4/ESFRI_SCRIPTA_VOL3_INNO_double_page.pdf

phylogeny-based assignment is also supported, based on RAxML-ng [39], EPA-ng [40], and Silva [37]. For the COI marker gene, the RDPClassifier [41] and the MIDORI database [42] are used for the taxonomic assignment. In addition, ecological and phylogenetic analysis are facilitated via the phyloseq R package [43].

All the pipeline- and third-party module–controlling parameters are defined in a plain "parameter-value pair" text file. Its straightforward format eases the analysis fine tuning, complementary to the aforementioned checkpoint mechanism. A tutorial about PEMA and installation guidance can be found on PEMA's GitHub repository[4].

2.1.4 Methods & Implementation

PEMA's architecture comprises 4 main parts taking place in tandem (Figure 2.1). A detailed description of the tools invoked by PEMA and their licenses is included in Additional File 1: Supplementary Methods.

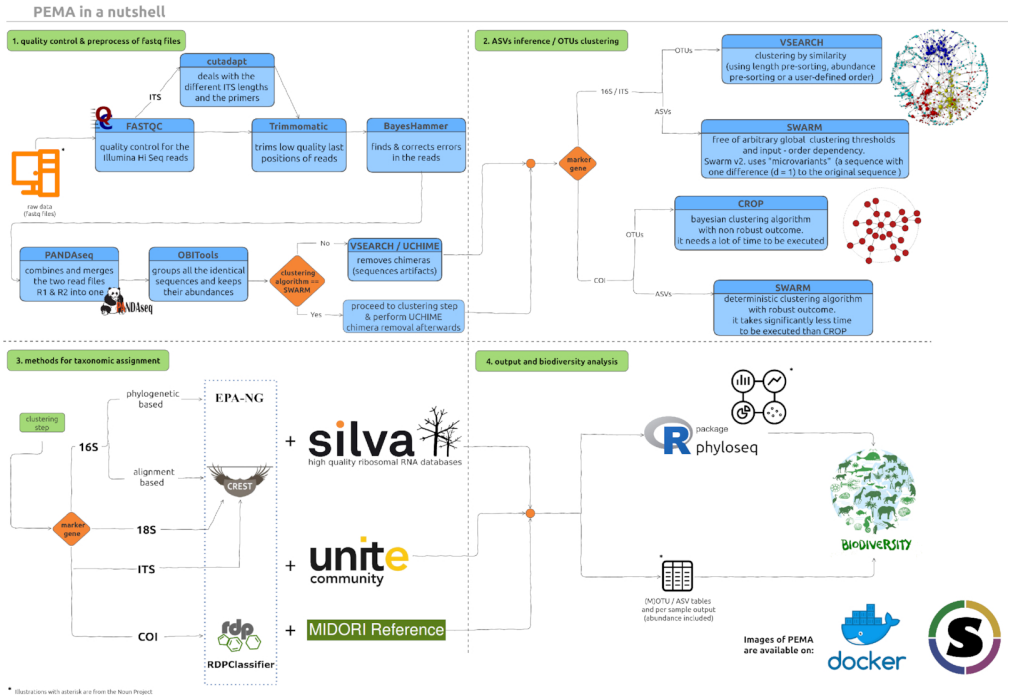

FIGURE 2.1: PEMA comprises 4 parts. The first step (top left) is the quality control and pre-processing of the Illumina sequencing reads. This step is common for both 16S rRNA and COI marker genes. The second step (top right) is the clustering of reads to (M)OTUs or their inferring to ASVs. The third step (bottom left) is the taxonomy assignment to the generated (M)OTUs/ASVs. In the fourth step (bottom right), the results of the metabarcoding analysis are provided to the user and visualized. *noun project icons by: ProSymbols (US), IconMark (PH), Nithinan Tatah (TH). clustering figure adapted from DOI: 10.7717/peerj.1420/fig-1

[4] https://github.com/hariszaf/pema

Part 1: Quality control and pre-processing of raw data

First, FastQC [44] is used to obtain an overall read-quality summary; visual inspection of each sample's quality may recommend removing those insufficient quality, as well as samples with a low number of reads, and rerunning the analysis. To correct errors produced by the sequencer, PEMA incorporates a number of tools. Trimmomatic [45] implements a series of trimming steps, which either remove parts of the sequences corresponding to the adapters or the primers, trim and crop parts of the reads, or even remove a read completely, when it fails to reach the quality-filtering standards set by the user. Cutadapt [46] is used additionally for the case of ITS to address the variability in length of this marker gene (see Additional File 1: Supplementary Methods). BayesHammer [47], an algorithm of the SPAdes assembly toolkit [48], revises incorrectly called bases. PANDAseq [49] assembles the overlapping paired-end reads, and then the obiuniq program of OBITools [50] groups all the identical sequences in every sample, keeping track of their abundances. The VSEARCH package [33] is then invoked for chimera removal; however, if the Swarm v2 algorithm is selected, this step will be performed after the ASV inference (see next section).

Part 2: (M)OTU clustering and ASV inference

Quality-controlled and processed sequences are subsequently clustered into (M)OTUs or treated as input for inferring ASVs. For the case of 16S and 18S rRNA marker genes, VSEARCH [33] is used for OTU clustering, while ASVs can be identified by the Swarm v2 algorithm [35]. VSEARCH is an accurate and fast tool that can handle large datasets; at the same time it is a great alternative for USEARCH [51] because it is distributed under an open source license.

For the ITS and COI marker genes, CROP [34], an unsupervised probabilistic Bayesian clustering algorithm that models the clustering process using birth-death Markov chain Monte Carlo (MCMC), is used. The CROP clustering algorithm is adjusted by a series of parameters that need to be tuned by the user (namely, b, e, and z). These parameters depend on specific dataset properties such as the length and the number of reads. PEMA automatically adjusts b, e, and z by collecting such information and applying the CROP recommended parameter-setting rules [34]. ASV inference is conducted by Swarm v2 [35] in this case too.

Because the Swarm v2 algorithm is not affected by chimeras (F. Mahé, personal communication), when Swarm v2 is selected, chimera removal occurs after the clustering (see Additional File 1: Supplementary Methods: Swarm v2). This leads to a computational time gain as chimeras are sought among ASVs, instead of ungrouped reads.

Last, any singletons, i.e., sequences with only 1 read, occurring after the (M)OTU clustering or the ASV inference may be removed according to the user's parameter settings.

Part 3: Taxonomy assignment

Alignment-based taxonomy assignment is supported for all marker gene analyses. In the case of the 16S/18S rRNA and ITS marker genes, the LCAClassifier algorithm of the CREST set of resources and tools [20] is used together with the Silva [37] and the Unite [52]

database, respectively, to assign taxonomy to the OTUs. Two versions of Silva are included in PEMA: 128 (29 September 2016) and 132 (13 December 2017). Because classifiers need first to be trained for each database they use, for future Silva [37] versions new PEMA versions will be available.

For the COI marker gene, PEMA uses the RDPClassifier [41] and the MIDORI reference database [42] to assign taxonomy of the MOTUs. The MIDORI database contains quality-controlled metazoan mitochondrial gene sequences from GenBank [53].

Intended primarily for studies from less explored environments, phylogeny - based assignment is available for 16S rRNA marker gene data. PEMA maps OTUs to a custom reference tree of 1,000 Silva-derived consensus sequences (created using RAxML-ng [39] and gappa [phat algorithm] [54], Figure 2.2A). PaPaRa [55] and EPA-ng [40] combine the OTU clustering output and the reference tree to produce a phylogeny-aware alignment and map the 16S rRNA OTUs to the custom reference tree. Beyond the context of PEMA, users may visualize the output with tree viewers such as iTOL [56] (Figure 2.2B).

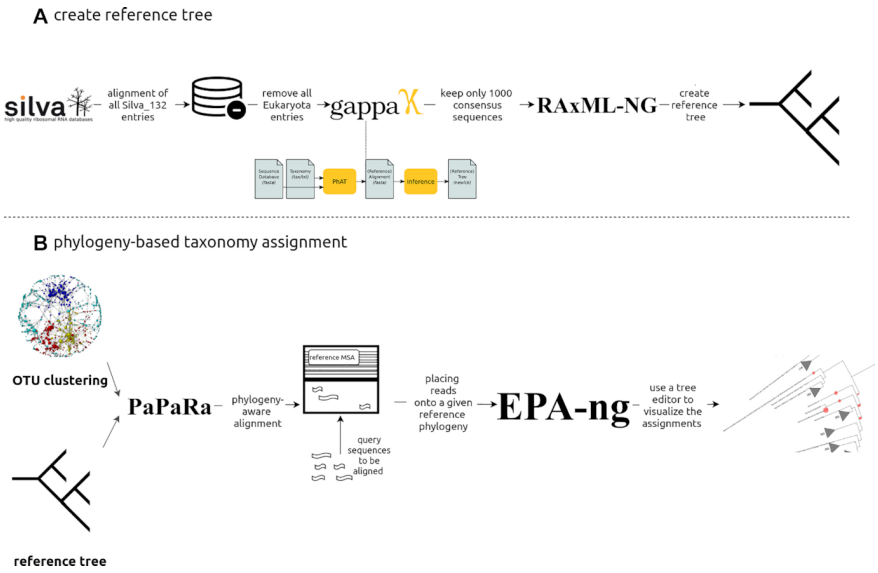

FIGURE 2.2: Phylogeny - based taxonomy assignment. A: Building a reference tree for the phylogeny-based taxonomy assignment to 16S rRNA marker gene OTUs: from the latest edition of Silva SSU, all entries referring to Bacteria and Archaea were used and using the "art" algorithm, 10,000 consensus taxa were kept. B: Using PaPaRa and the OTUs that come up from every analysis, an MSA was made and EPA - ng took over the phylogeny - based taxonomy assignment. *noun project icons by: Rockicon and A Beale.

Part 4: Ecological downstream analysis of the taxonomically assigned (M)OTU/ASV tables

PEMA's major output is either an (M)OTU or an ASV table with the assigned taxonomies and the abundances of each taxon in every sample. For each sample of the analysis, a

subfolder containing statistics about the quality of its reads, as well as the taxonomies and their abundances, is also returned.

Via the phyloseq R package [43], downstream ecological analysis of the taxonomically assigned OTUs or ASVs is supported. This includes $\alpha-$ and $\beta-$diversity analysis, taxonomic composition, statistical comparisons, and calculation of correlations between samples.

When selected, in addition to the phyloseq [43] output, a multiple sequence alignment (MSA) and a phylogenetic tree of the OTU/ASVs retrieved can be returned; for the MSA, the MAFFT [57, 58] aligner is invoked while the latter is built by RAxML-ng [39].

PEMA container-based installation

An easy way of installing PEMA is via its containers. A Dockerized PEMA version is available. Singularity users can *pull* the PEMA image from as described in PEMA GitHub repository. Between the 2 containers, the Singularity-based one is recommended for HPC environments owing to Singularity's improved security and file accessing properties, see here[5]. PEMA can also be found in the bio.tools (id: PEMA) and SciCruch (PEMA, RRID:SCR_017676) databases. For detailed documentation, see here[6].

PEMA output

All PEMA - related files (i.e., intermediate files, final output, checkpoint files, and per - analysis parameters) are grouped in distinct (self - explanatory) subfolders per major PEMA pipeline step. In the last subfolder, i.e., subfolder 8, the results are further split into folders per sample. This eases further analysis both within the PEMA framework (e.g., partial re-execution for parameter exploration) and beyond. An extra subfolder is created when an ecological analysis via the phyloseq package has been selected.

PEMA modules added after publication

7

2.1.5 Results & Validation

Evaluation

To evaluate PEMA, 2 approaches were followed. First, PEMA was benchmarked against mock community datasets. Second, PEMA was used to analyse previously published datasets. PEMA's output was then compared with the original study outcome, as well as with the output of QIIME2, LotuS, Mothur, and Barque (where applicable).

Four mock communities, 1 for each marker gene, were used. With respect to the 16S rRNA marker gene, a mock community of Gohl et al. [59] with 20 different bacterial species was studied. Correspondingly, in the case of the 18S rRNA marker gene, a mock

[5] https://dev.to/grokcode/singularity–a-docker-for-hpc-environments-i6p
[6] https://hariszaf.github.io/pema_documentation/
[7] REMEMBER TO WRITE THIS PART

community of Bradley et al. [60] with 12 algal species was used; for the ITS, one of Bakker [61] including 19 different fungal taxa; and for the case of the COI marker gene, a mock community of Bista et al. [62] containing 14 metazoan species. More information on the mock communities, their original studies, and the results of PEMA for various combinations of parameters can be found in Additional File 2: Mock Communities.

Complementary to the mock community evaluation, 2 publicly available datasets from published studies were investigated through PEMA. For the 16S rRNA marker gene, the dataset reported by Pavloudi et al. [63] was used; the original study aimed at investigating the sediment prokaryotic diversity along a transect river–lagoon–open sea. For the COI case, the dataset of Bista et al. [64] was used; this study investigated whether eDNA can be used for the accurate detection of chironomids (a taxonomic group of macroinvertebrates) in a freshwater habitat.

In both approaches, the respective .fastq files were downloaded from the European Nucleotide Archive (ENA) of the European Bioinformatics Institute ENA-(EBI) using *ENA File Downloader version 1.2* [65] and PEMA was run on the in-house HPC cluster.

All analyses were conducted on identical Dell M630 nodes (128 GB RAM, 20 physical Intel Xeon 2.60 GHz cores).

Mock community evaluation

PEMA was tested against mock communities. An evaluation of its accuracy must capture (i) how many of PEMA's predictions are true (i.e., the percent of correctly assigned taxa among all predicted taxa) and (ii) how many of the taxa existing in the mock community were recovered successfully by PEMA. The precision statistical metric was used to assess the former, and recall, the latter. In addition, the $F1$-score was used as a combined metric of both precision and recall. Precision is calculated as the ratio of true-positive results (TP) over the total number of true- (TP) and false-positive results (FP) predicted by a model, as follows: $precision = TP/(TP+FP)$; recall is the ratio of TP over the total number of TP and false-negative results (FN): $recall = TP/(TP+FN)$. The $F1$-score is the precision and recall harmonic mean and is calculated by means of the following formula: $F1 = 2 \times (precision \times recall)/(precision + recall)$ [66].

Adequate accuracy was achieved when PEMA was used to recover the marker gene–specific mock communities at the genus level. Precision and recall scores of ~80% or more were observed, with 2 exceptions in precision but also 3 very high scores in recall. Overall the F1-scores ranged from 74% to 86%. A detailed description of the benchmark methodology and statistics analysis is given in Additional File 2: Mock Communities.

Detailed presentation of per-marker-gene–specific mock community recovery via PEMA is provided in the following sections. Several different sets of parameters were chosen for each marker gene. Each marker gene has special features (e.g., length variability, sequence variability), and each Illumina run has its own intrinsic biases (e.g., primers used, PCR protocol); thus, parameter tuning plays a crucial part in metabarcoding analyses.

In an attempt to thoroughly analyse the sequence data from the mock communities, various sets of parameters were tested on the basis of the experimental details of the published studies but also in an exploratory way. Many different parameter settings were tested, especially for the steps of quality trimming of the reads and the OTU clustering/ASV

Marker gene	Precision	Recall	F1
16S rRNA	0.81	0.85	0.83
18S rRNA	0.75	0.90	0.82
ITS	0.79	0.94	0.86
COI	0.62	0.93	0.74

TABLE 2.1: Summary benchmark of PEMA marker - gene – specific mock community recovery (precision)

inference. The differences in their output indicate how sensitive this method is, as well as the great need of a mock community in every metabarcoding study—both as a control but also as a *tuning system* for the parameter setting of the pipeline used.

16S rRNA

When PEMA was performed with the Swarm v2 algorithm (d = 3, strictness = 0.6) without removal of singletons, 18 of the 20 taxa were identified to the genus level and 3 of these even to the species level. There were 2 species that were not found in any of the PEMA runs. According to Gohl et al. [59], there was a discrepancy in the identification of those 2 species that was dependent on the amplification protocol used. It is worth mentioning that as d increases, taxa cannot be identified to species level at all; however, FP assignments decrease. Thus, when d = 30 and strictness = 0.6 for the KAPA samples, Enterococcus was not identified at all; however, PEMA finds its greatest $F1$ value (at the genus level, see Table 2.1) as the FP assignments returned are minimized. When PEMA was run using the VSEARCH clustering algorithm, high precision values were returned in all cases (>0.79). However, the recall values were decreased when using Swarm v2 (0.65–0.68).

18S rRNA

When PEMA was performed using the Swarm v2 algorithm (d = 1, strictness = 0.5), 3 of 12 community members were identified to species level (*Isochrysis galbana*, *Nannochloropsis oculata*, and *Thalassiosira pseudonana*), 6 to genus, and the remaining 3 to class; the latter were all the green algae species (Chlorophyta) of the mock community. However, a better $F1$-score (0.82) was achieved when the class of Chlorophyceae was not found at all (d = 1, strictness = 0.3) because the FPs were decreased to only 1. When the VSEARCH algorithm was used, *I. galbana* was identified only to the genus level, the *Nannochloropsis* to the order level (Eustigmatales), and the *Poterioochromonas* genus to its class (Chrysophyceae).

ITS

When PEMA was performed using the Swarm v2 algorithm (d = 20) and targeting the ITS2 region, ASVs from 5 of the 19 species of the mock community were assigned to species level, 10 to genus, 2 to family, and 2 to class level. Contrary to the study by Bakker [61], PEMA identified the genus Chytriomyces in all 3 samples, as well as the Ustilaginaceae

family. Only 1 FP assignment was recorded. When the CROP algorithm was used, PEMA's output was less accurate; the *Fusarium* species contained in the mock community were not identified further than their family (Nectriaceae). As mentioned by Bakker [61], many reads deriving from the *Fusarium spp.* were not assigned to species level because of the quality-trimming step. In addition, a manually assembled reference database for the taxonomy assignment was used in the initial study, containing only sequences of the mock community species, which biased this step, making the results not directly comparable to our case.

COI

When PEMA was performed on the Bista et al. dataset [62] and using Swarm v2 (d = 10), it identified 12 of the 14 species included in the mock community. The sole non - identified species were *Bithynia leachii* and *Anisus vortex*. For *B. leachii* no entry exists in the MIDORI database, version MIDORI_LONGEST_1.1. However, the existence of another species of the genus *Bithynia* was recorded. With respect to *A. vortex*, PEMA returned a high abundance ASV assigned to the *Anisus* genus but with a low confidence level. PEMA managed to identify all the members of the mock community. This includes *Physa fontinalis*, which was originally not designed to be a member of the mock community but, as Bista et al. [62] explain, was recorded owing to cross - contamination. In the case of the COI marker gene, unique sequences with low abundances (singletons or doubletons) often lead to spurious MOTUs/ASVs. Thus, as shown in Additional File 2: Mock Communities, the FP assignments are decreased when these low-abundant sequences are removed; also, the abundance of the assignments (i.e., read counts) retrieved can indicate FP assignments. Thus, TP assignments occur in greater abundance, with hundreds or even thousands of reads—contrary to most of the FP results, whose abundance is < 10 read counts. That is mostly for the case of the COI marker gene because eukaryotes are under study; eukaryotes have a great number of copies of this marker gene — different numbers of copies among the different species — and not just a single one as is almost always the case in bacteria. Therefore, assignments with such low abundances should be doubted as TP results in analyses on real datasets.

Comparison with existing software

PEMA's features were compared with those of mothur [22], QIIME 2 [23], LotuS [24] and Barque. Table 2.2 presents a detailed comparison among the 4 tools' features in terms of marker gene support, diversity and phylogeny analysis capability, parameter setting and mode of execution, operation system availability, and HPC suitability. As shown, PEMA is equally feature - rich, if not richer in certain feature categories, compared with the other software packages. In particular, PEMA's support for COI marker gene studies is distinctive; 2 methods for taxonomy assignment are supported, and PEMA's easy parameter setting, step - by - step execution, and container distribution render it user and analysis friendly.

Feature	LotuS	QIIME 2	mothur	Barque	PEMA
16S rRNA	✓	✓	✓		✓
18S rRNA	✓	✓	✓		✓
ITS	✓	✓			✓
COI				✓	✓
diversity indices		✓	✓		✓
alignment-based taxonomy assignment	✓	✓	✓	✓	✓
phylogenetic-based taxonomy assignment	✓	✓			✓
parameters assigned in the command line	✓	✓	✓		
parameters assigned through a text file	✓			✓	✓
step-by-step execution	✓	✓	✓		✓
all steps in one go possible	✓			✓	✓
available for any Operating System (Linux, OSX, Windows)		✓	✓		✓
traditional application installation	✓	✓	✓	✓	
available as a virtual machine		✓			
available as a container		✓			✓
available for HPC as a container (Singularity container)					✓

TABLE 2.2: Comparison of the basic features of the different pipelines

Evaluation on real datasets and against other tools

In the following sections, a comparative study on real datasets of the 16S rRNA and COI marker genes is presented. Analyses using PEMA and the pipelines mentioned above that support each of these 2 marker genes were performed, both with multiple sets of parameters. It is typical for pipelines to invoke a variety of established tools. In many cases, a number of tools are common among different pipelines. Therefore, it is important to stress that such comparisons should not be taken into account strictly; declaring that one pipeline is better than another is not trivial. Potentials and limitations of both the pipelines and the metabarcoding method, as well as the importance of the role of the pipeline user, are underlined in the following sections.

16S rRNA marker gene analysis evaluation

To evaluate PEMA's performance, a comparative analysis of the Pavloudi et al. [63] dataset with mothur [22], QIIME 2 [23], LotuS [24] and PEMA was conducted.

		QIIME 2				
Parameter	**LotuS**	**mothur**	**Deblur**	**DADA2**	**PEMA**	**Pavloudi et al. [63]**
No. of OTUs	9,849	142,669	517	1,023	6,028	7,050
Execution time (h)	~9	~67[8]	2.5	~5	~1.5	~26

TABLE 2.3: OTU predictions and execution time for the different pipelines. * ~ 56 if the reference database is already built

It is known that the choice of parameters affects the output of each analysis; therefore, it is expected that different user choices might distort the derived outputs. For this reason and for a direct comparison of the pipelines, we have included all the commands and parameters chosen in the framework of this study in Additional File 1: Supplementary Methods. The results of the processing of the sequences by PEMA are presented in Table S1. All analyses were conducted on identical Dell M630 nodes (128 GB RAM, 20 physical Intel Xeon 2.60 GHz cores). LotuS, mothur, and QIIME 2 operated in a single-thread (core) fashion. PEMA, given the BDS intrinsic parallelization [27], operated with up to the maximum number of node cores (in this case 20).

The execution time and the reported OTU number of each tool are presented in Table 2.3. LotuS and PEMA resulted in a final number of OTUs comparable to that of Pavloudi et. al [63]. Clearly, owing to PEMA's parallel execution support, the analysis time can be significantly reduced (~ 1.5 hours in this case). The execution time depends on the parameters chosen for each software (see Additional File 1: Supplementary Methods).

Owing to the non - full overlap of the sequence reads, mothur resulted in an inflated number of OTUs; thus, it was excluded from further analyses. The results of all the pipelines were analysed with the phyloseq script that is provided with PEMA. The taxonomic assignment of the PEMA - retrieved OTUs is shown in Figure 2.3. The phyla that were found in the samples are similar to the ones that were found in the original study [63]. Although the lowest number of OTUs was found in the marine station (Kal) (Supplementary Table S3), which is not in accordance with Pavloudi et. al [63], the general trend of a decreasing number of OTUs with increasing salinity was observed as in the original study (Supplementary Figure S1). Notably, this result was not observed with the other tested pipelines (Supplementary Table S3). Furthermore, each of the pipelines resulted in a different taxonomic profile (Supplementary Figures S2–S4), with an extreme case of missing the order of Betaproteobacteriales (Supplementary Figures S5–S7).

Moreover, when the PERMANOVA analysis was run for the results of PEMA, LotuS, and DADA2, it was clear that the microbial community composition was significantly different in each of the 3 sampled habitats (i.e., river, lagoon, open sea) (PERMANOVA: F.Model = 7.0718, $P < 0.001$; F.Model = 6.5901, $P < 0.001$; F.Model = 2.2484, $P < 0.05$, respectively), which is in accordance with Pavloudi et al. [63]. However, this was not the case with Deblur (PERMANOVA: $P > 0.05$). Overall, PEMA's output is in accordance with the original study [63], and seen through this perspective PEMA performed equally well with the other tested pipelines, along with having the shortest execution time.

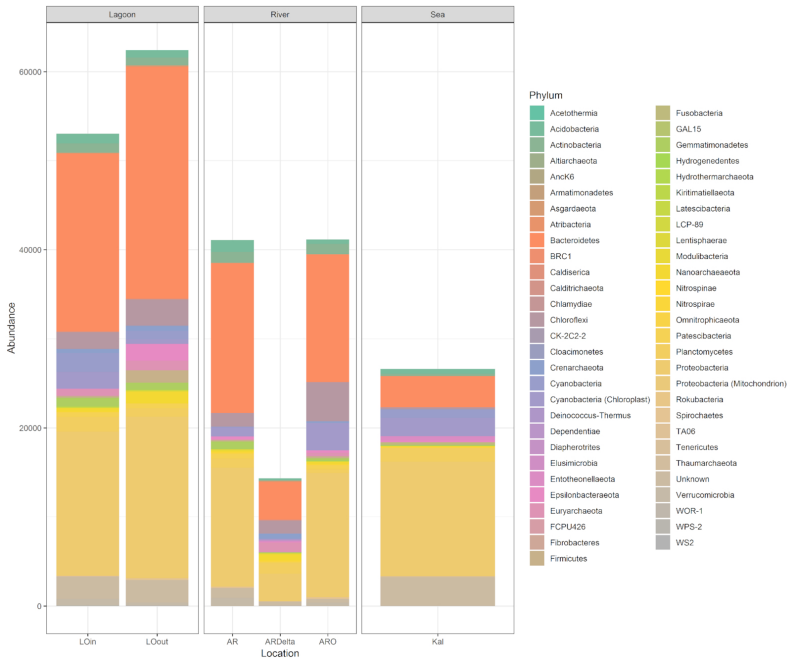

FIGURE 2.3: OTU bar plot at the phylum level. Bar plot depicting the taxonomy of the retrieved OTUs from PEMA for the dataset of Pavloudi et al. [63], at the phylum level for the case of the 16S marker gene. AR: Arachthos; ARO: Arachthos Neochori; ARDelta: Arachthos Delta; LOin: Logarou station inside the lagoon; LOout: Logarou station in the channel connecting the lagoon to the gulf; Kal: Kalamitsi.

COI marker gene analysis evaluation

Bista et al. [64] created 2 COI libraries of different sizes: COIS (235 - bp amplicon size) and COIF (658 - bp amplicon size). The sequencing reads of COIS were selected for PEMA's evaluation; the COIF sequencing read pairs had no overlap so as to be merged and therefore were not considered appropriate for the analysis.

As previously, PEMA's performance was evaluated through a comparative analysis of the Bista et al. [64] dataset with Barque[9]; the commands and parameters chosen can be found in Additional File 1: Supplementary Methods. Regarding the creation of the MOTU table, in the Bista et al. [64] study VSEARCH [33] was used with a clustering at 97% similarity threshold. Afterwards, the BLAST+ (megablast) algorithm [67] was used against a manually created database including all NCBI GenBank COI sequences of length > 100 bp (June 2015) while excluding environmental sequences and higher taxonomic level information [64]. As discussed in the publication, this approach resulted in 138 unique MOTUs of which 73 were assigned to species level. For PEMA's evaluation, the chosen clustering algorithm was Swarm v2, using different options for the cluster radius (d) parameter (Table 2.4); according to Mahé et al. [35], this is the most important parameter because it affects the number of MOTUs that are being created. The resulting MOTUs

[9] https://github.com/enormandeau/barque

Parameter	***d* = 1**	***d* = 2**	***d* = 3**	***d* = 10**	***d* = 13**
MOTUs after pre-process and clustering steps	83,791	59,833	33,227	7,384	4,829
MOTUs after chimera removal	80,347	57,863	32,539	7,339	4,796
Non-singleton MOTUs	6,381	4,947	2,658	1,914	1,634
Assigned species	62	83	86	86	84
Execution time (h)	2:01:35	2:09:49	1:51:44	2:17:26	2:31:15

TABLE 2.4: PEMA'sa output and execution time; PEMA's output and execution time (using a 20-core node) for different values of Swarm's d parameter.

were classified against the MIDORI reference database [42] using RDPClassifier [41]. The results of the processing of the sequences are reported in Supplementary Table S3. For the case of Barque, the BOLD Database was used [68].

As shown in Table 2.4, PEMA resulted in 83 species-level MOTUs with a cluster radius (d) of 2, which is similar to the findings of the published study (i.e., 73 species). Although both the clustering algorithm and the taxonomy assignment methods were different between the original [64] and the present study, the results regarding the number of unique species present in the samples are in agreement to a considerable extent.

The computational time required by PEMA for the completion of the analysis is also reported in Table 2.4. Regardless of the value of the d parameter, all analyses were completed in ~ 2 hours, i.e., fast enough to allow parameter testing and customization. Regarding Barque, the analysis resulted in the identification of 51 species-level MOTUs and was concluded in 15 minutes. This difference is due to the error correction step of PEMA (BayesHammer algorithm [47]), which plays an important part in the enhanced results that PEMA returns, but it also requires a certain computational time; Barque does not have an analogous step, and therefore its overall execution time is shorter.

PEMA performed better than Barque at identifying taxa that were included in the positive control contents of the published study (Table 2.5).

2.1.6 Discussion

OTU clustering vs ASV inference

There is an ongoing discussion about whether ASVs exceed OTUs. The strongest argument to this end is that ASVs are real biological sequences. Hence, they can be compared between different studies in a straightforward way; considered as consistent labels. In comparison, de novo OTUs are constructed, or "clustered," with respect to the emergent features of each specific dataset. Therefore, OTUs defined in 2 different datasets cannot be directly compared.

However, the OTU concept is not compulsorily related to the clustering approach; it is widely used to describe results based on its biological meaning but it does not imply clustering. In addition, according to Callahan et al. [31], "ASV methods infer the biological

Barque	**PEMA**	**Bista et al. [50]**
*Ablabesmyia monilis**	*Ablabesmyia monilis**	*Ablabesmyia monilis*
	*Crangonyx pseudogracilis**	*Crangonyx pseudogracilis*
	*Radix sp.**	*Radix sp.*
	Chironomidae sp.*	Chironomidae sp.
	*Ancylus sp.***	*Ancylus fluviatilis*
	*Athripsodes aterrimus, Athripsodes cinereus***	*Athripsodes albifrons*
*Chironomus anthracinus***	*Chironomus sp., Chironomus anthracinus, Chironomus pseudothummi, Chironomus riparius***	*Chironomus tentans*
*Polypedilum sordens***		*Polypedilum nubeculosum*
*Athripsodes aterrimus***		*Athripsodes albifrons*

TABLE 2.5: Comparison of the taxonomy of retrieved MOTUs among PEMA, Barque, and the positive controls of Bista et al. [64] ; * Taxonomies identical to the published study (species level), ** Taxonomies identical to the published study (genus level).

sequences in the sample prior to the introduction of amplification and sequencing errors, and distinguish sequence variants differing by as little as one nucleotide." As a result, ASVs could be considered as OTUs of higher resolution.

It is due to this concept confusion that algorithms whose rationale is considerably closer to the variant-based approach are still considered as OTU clustering algorithms [31]. Swarm v2 produces all possible *microvariants* of an amplicon to implement an exact-string comparison [35]. Furthermore, real biological sequences, *clouds of microvariants* are produced as its output, which can be used for comparisons between different studies. Thus, Swarm v2 can be considered as an ASV-inferring algorithm.

Traditional clustering methods have certain limitations such as arbitrary global clustering thresholds and centroid selection because they depend on the input order and are time-consuming, etc. [69], which variant-based approaches manage to address. However certain algorithms for OTU clustering such as VSEARCH have been proven to be especially reliable, and they are widely used by many researchers. Furthermore, ASVs intend to improve taxonomic resolution; however, a vast number of inferred ASVs (see here for more) can lead to inflation of diversity estimates, especially in the case of microbial communities, thus making the analysis even more complicated.

ASV or OTU approaches are supported by PEMA, although we have found that similar ecological results are produced by both these methods, as also suggested by Glassman and Martiny [70].

Beyond environmental ecology, ongoing and future work

PEMA is mainly intended to support eDNA metabarcoding analysis and be directly applicable to next - generation biodiversity / ecological assessment studies. Given that

community composition analysis may also serve additional research fields, e.g., microbial pathology, the potential impact of such pipelines is expected to be much higher. Ongoing PEMA work focuses on serving a wide scientific audience and on making it applicable to more types of studies. The easy set - up and execution of PEMA allows users to work closely with national and European HPC / e - infrastructures (e.g., ELIXIR Greece, LifeWatch ERIC, EMBRC ERIC). To that end and in a mid - term perspective, a CWL version of PEMA will be explored. The aim of this effort is to reach out to a wider scientific audience and address both their ongoing as well as future analysis needs.

By supporting the analysis of the most commonly used marker genes for Bacteria and Archaea (16S rRNA), Fungi (ITS), and Metazoa (COI/18S rRNA), a holistic biodiversity assessment approach is now possible through PEMA and eDNA metabarcoding; although, from a mid-term perspective, it is our intention to allow ad hoc and in - house databases to be used as reference for the taxonomy assignment.

Conclusions

PEMA is an accurate, execution - friendly and fast pipeline for eDNA metabarcoding analysis. It provides a per - sample analysis output, different taxonomy assignment methods, and graphics - based biodiversity / ecological analysis. This way, in addition to (M)OTU/ASV calling, it provides users with both an informative study overview and detailed result snapshots.

Thanks to a nominal number of installation and execution commands required for PEMA to be set and run, it is considered essentially user friendly. In addition, PEMA's strategic choice of a single parameter file, implementation programming language, and multiple container - type distribution grant it speed (running in parallel), on - demand partial pipeline enactment, and provision for HPC - system – based sharing.

All the aforementioned features render PEMA attractive for biodiversity / ecological assessment analyses. By supporting the analysis of the most commonly used marker genes for Prokaryotes (Bacteria and Archaea), as well as Eukaryotes (Fungi and Metazoa), PEMA allows assessment of biodiversity in different levels of biodiversity. Applications may mainly concern environmental ecology, with possible extensions to such fields as microbial pathology and gut microbiome, in line with modern research needs, from low volume to big data.

2.1.7 Supplementary Material

You may find the Supplementary files of this study through PEMA's publication [10]

Additional File 1: Supplementary Methods: Description of tools invoked by PEMA and their licences. Description of the commands, along with their parameters, used to run PEMA, mothur, LotuS, and QIIME 2.

Additional File 2: Mock Communities: Details about the mock communities chosen and their corresponding studies, as well as the returned output of PEMA for each for a number of sets of parameters.

[10] https://academic.oup.com/gigascience/article/9/3/giaa022/5803335#supplementary-data

Supplementary Table S1: Number of sequences after each pre-processing step for the case of 16S rRNA gene.

Supplementary Table S3: Number of sequences after each pre-processing step for the case of COI, dataset from Bista et al. [64].

Supplementary Table S2: Diversity indices of the samples.

Supplementary Figure S1: Linear regression between the number of OTUs (averaged per sampling station) and the salinity of the sampling stations. L: Lagoon; S: Sea; R: River; AR: Arachthos; ARO: Arachthos Neochori; ARDelta: Arachthos Delta; LOin: Logarou station inside the lagoon; LOout: Logarou station in the channel connecting the lagoon to the gulf; Kal: Kalamitsi.

Supplementary Figure S2: Bar plot depicting the taxonomy of the retrieved OTUs from LotuS at the phylum level.

Supplementary Figure S3: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using Deblur at the phylum level.

Supplementary Figure S4: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using DADA2 at the phylum level.

Supplementary Figure S5: Bar plot depicting the taxonomy of the retrieved OTUs from LotuS at the class of Betaproteobacteriales.

Supplementary Figure S6: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using Deblur at the class of Betaproteobacteriales.

Supplementary Figure S7: Bar plot depicting the taxonomy of the retrieved OTUs from PEMA at the class of Betaproteobacteriales.

2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data[11]

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

2.2.1 Abstract

The mitochondrial cytochrome C oxidase subunit I gene (COI) is commonly used in environmental DNA (eDNA) metabarcoding studies, especially for assessing metazoan diversity. Yet, a great number of COI operational taxonomic units (OTUs) or/and amplicon sequence variants (ASVs) retrieved from such studies do not get a taxonomic assignment with a reference sequence. To assess and investigate such sequences, we have developed the Dark mAtteR iNvestigator (DARN) software tool. For this purpose, a reference COI-oriented phylogenetic tree was built from 1,593 consensus sequences covering all the three domains of life. With respect to eukaryotes, consensus sequences at the family level were constructed from 183,330 sequences retrieved from the Midori reference 2 database, which represented 70% of the initial number of reference sequences. Similarly, sequences from 431 bacterial and 15 archaeal taxa at the family level (29% and 1% of the initial number of reference sequences respectively) were retrieved from the BOLD and the PFam databases. DARN makes use of this phylogenetic tree to investigate COI pre-processed sequences of amplicon samples to provide both a tabular and a graphical overview of their phylogenetic assignments. To evaluate DARN, both environmental and bulk metabarcoding samples from different aquatic environments using various primer sets were analysed. We demonstrate that a large proportion of non-target prokaryotic organisms, such as bacteria and archaea, are also amplified in eDNA samples and we suggest prokaryotic COI sequences to be included in the reference databases used for the taxonomy assignment to allow for further analyses of dark matter. DARN source code is available on GitHub at https://github.com/hariszaf/darn and as a Docker image at https://hub.docker.com/r/hariszaf/darn.

2.2.2 Introduction

Metabarcoding: concept and caveats

DNA metabarcoding is a rapidly evolving method that is being more frequently employed i n a range of fields, such as biodiversity, biomonitoring, molecular ecology and others [11, 12]. Environmental DNA (eDNA) metabarcoding, targeting DNA directly isolated from environmental samples (e.g., water, soil or sediment, [13]), is considered a holistic approach (Stat et al. 2017) in terms of biodiversity assessment, providing high detection

[11] For author contributions and supplementary material please refer to the relevant sections. Modified version of the published review.

capacity. At the same time, it allows wide-scale rapid bio-assessment [14] at a relatively low cost as compared to traditional biodiversity survey methods [15].

The underlying idea of the method is to take advantage of genetic markers, i.e. marker loci, using primers anchored in conserved regions. These universal markers should have enough sequence variability to allow distinction among related taxa and be flanked by conserved regions allowing for universal or semi-universal primer design [16]. In the case of eukaryotes, the target is most commonly mitochondrial due to higher copy numbers than nuclear DNA and the potential for species level identification. Furthermore, mitochondria are nearly universally present in eukaryotic organisms, especially in case of metazoa, and can be easily sequenced and used for identification of the species composition of a sample [71]. However, it is essential that comprehensive public databases containing well curated, up-to-date sequences from voucher specimens are available [72]. This way, sequences generated by universal primers can be compared with the ones in reference databases, assessing sample OTU composition. The taxonomy assignment step of the eDNA metabarcoding method and thus, the identification via DNA-barcoding, is only as good and accurate as the reference databases [73].

Nevertheless, there is not a truly "universal" genetic marker that is capable of being amplified for all species across different taxa [74]. Different markers have been used for different taxonomic groups [11]. While bacterial and archaeal diversity is often based on the 16S rRNA gene, for eukaryotes a diverse set of loci is used from the analogous eukaryotic rRNA gene array (e.g., ITS, 18S or 28S rRNA), chloroplast genes (for plants) and mitochondrial DNA (for eukaryotes) in an attempt for species - specific resolution [30]. The mitochondrial cytochrome c oxidase subunit I (COI) marker gene has been widely used for the barcoding of the Animalia kingdom for almost two decades [75]. There are cases where COI has been the standard marker for metabarcoding, such as in the assessment of freshwater macroinvertebrates [76] even though not all taxonomic groups can be differentiated to the species level using this locus [11]; for example, in case of fish other loci are widely used such as 12S rRNA gene (hereafter referred to as 12S rRNA) [77].

The COI locus

The mitochondrial cytochrome c oxidase subunit I (also called cox1 or/and COI) is a gene fragment of 700 bp, widely used for metazoan diversity assessment. Here we present some of the reasons that microbial eukaryotes and prokaryotes are also amplified in such studies, raising the issue of the known unknown sequences. COI is a fundamental part of the heme aa3-type mitochondrial cytochrome c oxidase complex: the terminal electron acceptor in the respiratory chain. Even if aa3-type Cox have been found in bacteria, there are also other cytochrome c oxidase (Cox) groups, such as the cbb3-type cytochrome c oxidases (cbb3-Cox) and the cytochrome ba3 [78, 79].

Furthermore, the presence of highly divergent nuclear mitochondrial pseudogenes (numts) has been a widely known issue on the use of COI in barcoding and metabarcoding studies, leading to overestimates of the number of taxa present in a sample [80]. Numts are nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms [81].

Thus, as Mioduchowska et al. (2018) [82] highlight, when universal primers are used targeting the COI locus, it is possible to co-amplify both non-target numts and prokaryotes [83]. This has led to multiple erroneous DNA barcoding cases and it is now not rare to encounter bacterial sequences described as metazoan in databases such as GenBank [82].

Even though there are various known issues [16], COI is indeed considered as the "gold standard" for community DNA metabarcoding of bulk metazoan samples [84]; bulk is an environmental sample containing mainly organisms from the taxonomic group under study providing high quality and quantity of DNA [85]. However, as highlighted in the same study, this is not the case for eDNA samples. As Stat et al. (2017) [14] state, in the case of eDNA samples, the target region for metazoa is found in general at considerably lower concentrations compared to those from prokaryotes because most primers targeting the COI region amplify large proportions of prokaryotes at the same time [86, 87, 88]. Cold-adapted marine gammaproteobacteria are an indicative example for this case as shown by Siddall et al. (2009) [83].

2.2.3 Contribution

The co-amplification of prokaryotes explained above, is a major reason for why many Operational Taxonomic Units (OTUs) and/or Amplicon Sequence Variants (ASVs) in eDNA metabarcoding studies cannot get taxonomy assignments when metazoan reference databases are used (c.f. Aylagas et al. 2016 [89]) or they are assigned to metazoan taxa but with very low confidence estimates. Despite the presence of such OTUs/ASVs to a varying degree in metabarcoding studies using the COI marker gene [83], to the best of our knowledge, there has not been a thorough investigation of the origin for these sequences. Although unassignable sequences could be informative, there have been few attempts to further investigate this dark matter (e.g., [90, 91]).

The aim of this study was to build a framework for extracting such non-target, potentially unassigned (or assigned with low confidence) sequences from COI environmental sequence samples, hereafter referred to as "dark matter" as per Bernard et al. (2018) [92]. We argue that the vast majority of these sequences represent microbial taxa, such as bacteria and archaea.

More specifically, based on the previously described methodology by Barbera et al. (2019) [40] (see also full stack example of the EPA-ng algorithm) for large-scale phylogenetic placements, we built a framework to estimate to what extent the OTUs/ASVs retrieved in an environmental sample represent target taxa or not. That is, to evaluate the taxonomy assignment step in a metabarcoding analysis, by checking the phylogenetic placement of dark matter sequences. Similar studies have provided great insight into other marker genes, e.g. [93].

2.2.4 Methods & Implementation

Building the COI tree of life

Sequences for the COI region from all the three domains of life were retrieved from curated databases. Eukaryotic sequences were retrieved from the Midori reference 2 database (version: GB239) [42]. Initially, 1,315,378 sequences were retrieved corresponding to

183,330 unique species from all eukaryotic taxa. With respect to bacteria and archaea, 3,917 bacterial COI sequences were obtained from the BOLD database [68]. Similarly, 117 sequences from archaea were obtained from BOLD. In addition, for all the PFam protein sequences related to the accession number for COX1 (PF00115), the respective DNA sequences were extracted from their corresponding genomes. This way an additional 217 archaeal and 9,154 bacterial sequences were obtained (see Table 1). In total, sequences from 15 archaeal, 371 bacterial families and 60 taxonomic groups of higher level not assigned in the family level, were gathered. An overview of the approach that was followed is presented in Figure 2.4.

The large number of obtained sequences effectively prevents a phylogenetic tree construction encompassing their total number in terms of building a single phylogenetic tree covering all of the three domains of life (archaea, bacteria, eukaryota). Therefore, consensus representative sequences from each of the three datasets were constructed using the PhAT algorithm [54]; based on the entropy of a set of sequences, PhAT groups sequences into a given target number of groups so they reflect the diversity of all the sequences in the dataset. As PhAT uses a multiple sequence alignment (MSA) as input, all the three domain-specific datasets were aligned using the MAFFT alignment software tool v7.453 [57, 58].

Resources	**bacteria**		**archaea**	
	# of sequences	**# of strains**	**# of sequences**	**# of strains**
BOLD	3,917	2,267	117	117
PFam-oriented	9,154	4,532	217	115

TABLE 2.6: Number of sequences and taxonomic species per domain of life and resources. The (#) symbols stands for "number".

In the case of Eukaryotes, the alignment of the corresponding sequences would be impractically long because of their large number (183K sequences). To address this challenge, a two-step procedure was followed; a sequence subset of 500 sequences (*reference set*) was selected and aligned and then used as a backbone for the alignment of all the remaining eukaryotic COI sequences. All sequences were considered reliable as they were retrieved from curated databases (Midori2 and BOLD). To build the reference set, a number (n) of the longest sequences from each of the various phyla were chosen, proportionally to the number (m) of sequences of each phylum (see Supplementary Table 2.6). The –min-tax-level parameter of the PhAT algorithm corresponded to the class level, for the case of eukaryotes and to the family level for archaea and bacteria. This parameter forced the PhAT algorithm to build at least one consensus sequence for each class and family respectively. The taxonomy level was not the same for the case of eukaryotes sequence dataset and those of bacteria and archaea, as the number of unique eukaryotic families was one order of magnitude higher. The PhAT algorithm was invoked through the gappa v0.6.1 collection of algorithms [94].

A total of 1,109 consensus sequences (70% of total consensus sequences) were built covering the eukaryotic taxa, while 463 (29%) bacterial and 21 (1%) archaeal consen-

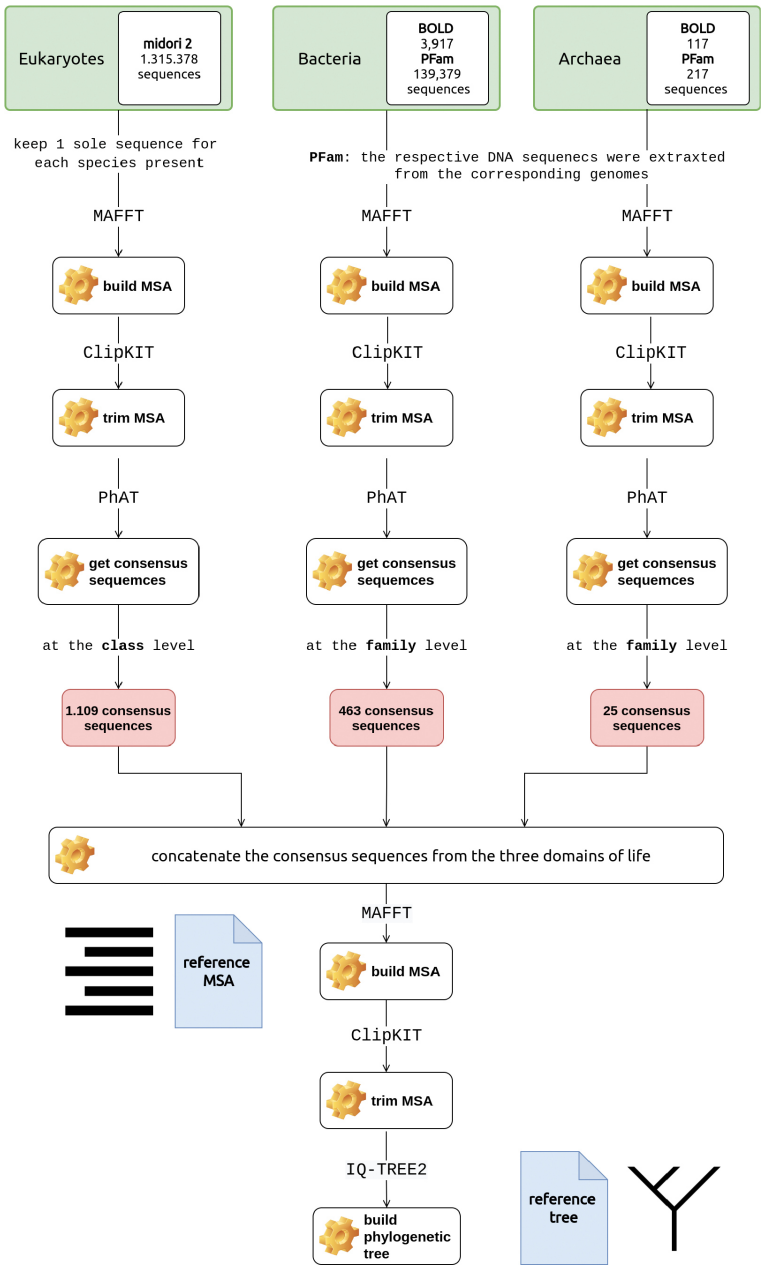

FIGURE 2.4: Overview of the approach followed to build the COI reference tree of life. Sequences were retrieved from Midori 2 (eukaryotes) and BOLD (bacteria and archaea) repositories. Consensus sequences at the family level were built for each domain specific dataset. MAFFT and consensus sequences at the family level were built using the PhAT algorithm. The COI reference tree was finally built using IQ-TREE2. Noun project icons by Arthur Slain and A. Beale.

sus sequences were included. The per-domain, consensus sequences returned can be found under the consensus_seqs directory on the GitHub repository (see _consensus.fasta files). These sequences were then merged as a single dataset and aligned to build a reference MSA; this time MAFFT was set to return using the –globalpair algorithm and the –maxiterate parameter equal to 1,000. The MSA returned was then trimmed with the ClipKIT software package [95] to keep only phylogenetically informative sites. The final MSA is available on GitHub, see trimmed_all_consensus_aligned_adjust_dir.aln.

The reference tree was then built based on this trimmed MSA using the IQ-TREE2 software [96, 97]. ModelFinder was invoked through IQ-TREE2 and the GTR+F+R10 model was chosen based on the Bayesian Information Criterion (BIC) among 286 models that were tested. The phylogenetic tree was then built using 1,000 bootstrap replicates (-B 1,000) and 1,000 bootstrap replicates for Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) (1,000 1000).

In the .iqtree file there are the branch support values; SH-aLRT support (%) / ultrafast bootstrap support (%).

A thorough description of all the implementation steps for building the reference tree is presented in this Google Collab Notebook. The computational resources of the IMBBC High Performance Computing system, called Zorba [98], were exploited to address the needs of the tasks.

Investigating COI dark matter

The COI reference tree was subsequently used to build and implement the Dark mAtteR iNvestigator (DARN) software tool. DARN uses a .fasta file with DNA sequences as input and returns an overview of sequence assignments per domain (eukaryotes, bacteria, archaea) after placing the query sequences of the sample on the branches of the reference tree. Sequences that are not assigned to a domain are grouped as "distant". It is necessary for the input sequences to represent the proper strand of the locus, i.e. input reads should have forward orientation. Optionally, DARN invokes the orient module of the vsearch package [33] to implement this step, in case the user is not sure about the orientation of the sequences to be analysed.

The focal query sequences are aligned with respect to the reference MSA using the PaPaRa 2.0 algorithm [55]. The query sequences are then split to build a discrete query MSA. Finally, the Evolutionary Placement Algorithm EPA-ng [40] is used to assign the query sequences to the reference tree.

To visualise the query sequence assignments, a two-step method was developed. First, DARN invokes the gappa examine assign tool which taxonomically assigns placed query sequences by making use of the likelihood weight ratio (LWR) that was assigned to this exact taxonomic path. In the DARN framework, by making use of the –per-query-results and –best-hit flags, the gappa assign software assigns the LWR of each placement of the query sequences to a taxonomic rank that was built based on the taxonomies included in the reference tree. The first flag ensures that the gappa assign tool will return a tabular file containing one assignment profile per input query while the latter will only return the assignment with the highest LWR. DARN automatically parses this output of gappa assign to build two input Krona profile files based on

- the LWR values of each query sequence and
- an adjustive approach where all the best hits get the same value in a binary approach (presence - absence)

In the final_outcome directory that DARN creates, two .html files, one for each of the Krona plots; Krona plots are built using the ktImportText command of KronaTools [99]. In addition four .fasta files are generated including the sequences of the sample that have been assigned to each domain or as "distant". A .json file with the metadata of the analysis is also returned including the identities of the sequences assigned to each domain.

DARN also runs the gappa assign tool with the –per-query-results flag only. This way, the user can have a thorough overview of each sample's sequence assignments, as a sequence may be assigned to more than one branch of the reference tree, sometimes even to different domains. However, in cases with sequences assigned to multiple branches, the likelihood scores are most typically up to 100-fold to 1000-fold different.

DARN source code as well as all data sequences and scripts for building the reference phylogenetic tree are available on GitHub.

2.2.5 Results & Validation

Evaluation of the phylogenetic tree

The inferred phylogenetic tree is shown in Figure 2.5, with the bacterial (light blue) and archaeal (dark green) branches highlighted; in Supplementary material 3: Figure S1 the distribution of the eukaryotic phyla on the tree is presented. As shown, bacteria and archaea can be distinguished from eukaryotes. Scattered bacterial branches that are present among eukaryotic ones represent the diversity of the COI locus. To evaluate the phylogenetic tree, the set of consensus sequences were placed on it using the EPA-ng algorithm. The placements (see .jplace through a phylogenetic tree viewer, e.g. iTOL) verified that the phylogenetic tree built is valid, as the consensus sequences have been placed in their corresponding taxonomic branches (Supplementary material 4: Figure S2; the figure was built using the heat-tree module of the gappa examine tool).

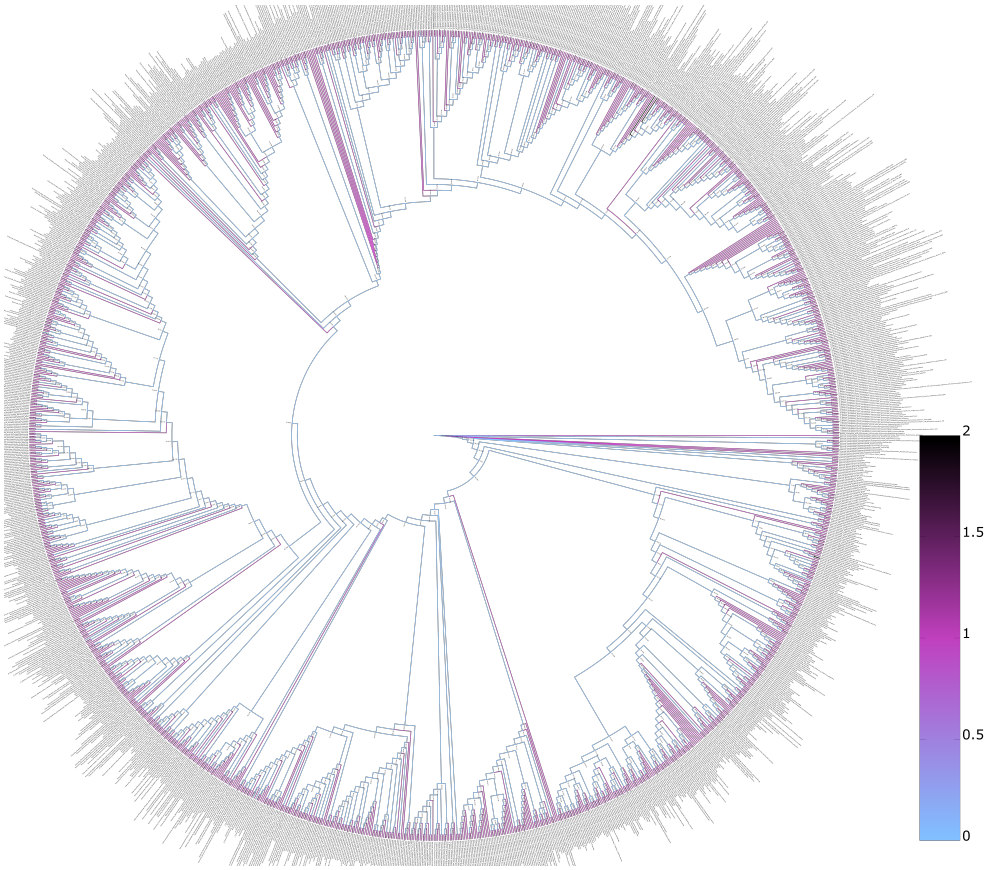

FIGURE 2.5: Phylogenetic tree of the consensus sequences retrieved; the tree that DARN makes use of. Light blue: bacterial branches. Dark green: archaeal branches. White: eukaryotic branches.

DARN using mock community data

To examine whether the phylogenetic-based taxonomy assignment addresses a real-world issue, a local blast database was built using the total number of the consensus sequences retrieved. As expected, when the consensus sequences were blasted against this local blastdb, all were matched with their corresponding sequences. However, when a mock dataset was used to evaluate the two approaches (blastdb and the phylogenetic tree) none of the bacterial sequences were captured as bacteria after blastn against the local blastdb (see output file here). All bacterial sequences returned an incorrect eukaryotic assignment. Contrarily, when the phylogenetic tree was used, all the bacterial sequences were captured.

DARN using real community data

To evaluate DARN on the presence of dark matter we analysed a wide range of cases to show the ability of DARN to detect and estimate dark matter under various conditions.

Both eDNA and bulk samples, from marine, lotic and lentic environments, were selected to reflect various combinations of primer and amplicon lengths, PCR protocols and bioinformatics analyses (Table 2.7).

More specifically, 57 marine, surface water, eDNA samples from Ireland were analysed through a. QIIME2 [23] and DADA2 [100] and, b. PEMA [101]. Similarly, 18 mangrove and 18 reef marine eDNA samples from Honduras, were analyzed using a. JAMP v0.74 and DnoisE [102] and b. PEMA Furthermore, a sediment sample and two samples from Autonomous Reef Monitoring Structures (ARMS) one conserved in DMSO and another in ethanol from the Obst et al. (2020) [103] dataset were analysed using PEMA. In addition, one lotic and two lentic samples from Norway were analysed using PEMA. For the case of the lentic samples, multiple parameter sets regarding the ASVs inference step were implemented; i.e the d parameter of the Swarm v2 [35] that PEMA invokes was set equal to 2 and 10 to cover a great range of different cases [104]. DARN was then executed using the ASVs retrieved in each case as input. All the DARN analyses and the PEMA runs were performed on an Intel(R) Xeon(R) CPU E5649 @ 2.53GHz server of 24 CPUs and 142 GB RAM in the Area52 Research Group at the University College Dublin.

The number of sequences returned, using various bioinformatic analyses, ranged from circa 3k to 214k (Table 2.7) in the different amplicon datasets used. A coherent visual representation of the DARN outcome for all the datasets is available here[12]. The visual and interactive properties of the Krona plot allow the user to navigate through the taxonomy. Furthermore, DARN also supports a thorough investigation per OTU/ASV, as it returns a .json file with all the OTUs/ASVs ids that have been assigned in each of the four categories (Bacteria, Archaea, Eukaryotes and distant).

Significant proportions of non-eukaryote DARN assignments were observed in all marine eDNA samples (Table 2.7). Bacterial assignments made up the largest proportion of the non-eukaryotic assignments (35.3% on average and more than 75% of the OTUs/ASVs in some cases), however, archaeal assignments were also detected to a great extent as well (18.4% on average). The lentic samples were those with the shortest amplicon length among those analysed (142 bp); hence, for their orientation a database with only the shortest consensus sequences (< 700 bp) was used, as otherwise a great number of sequences did not have sufficient number of hits and was discarded (see Suppl. material 2: Table S2). It is worth mentioning that in this case, the initial number of raw reads ranged from 53,000 (ERS6488992, ERS6488993) to 88,000 (ERS6488993) while the number of ASVs returned (using Swarm with d parameter equal to 10) ranged from 365 (ERS6488993) to 823 (ERS6488993). This relatively low number of ASVs could indicate that targeting such small COI regions could decrease the co-amplification of non-targeted sequences. In the case of bulk samples (Table 2.7) only a low proportion of the sequences were not assigned as Eukaryotes, suggesting that non-eukaryotic sequences are more abundant in environmental samples. This could be expected since prokaryotes are amplified as whole organisms from environmental samples, while metazoa that are usually the targeted taxa in COI studies, are amplified from DNA traces or/and other parts of biological source material.

[12] https://hariszaf.github.io/darn/

Sample(s) accession number	Sample type	Primer set	Amplicon length (bp)	Bioinfo pipeline(s)	# of ASVs	~% of sequence assignments per domain (if PEMA[a])			
						Eukaryotes	Bacteria	Archaea	distant
ERS6449795-ERS6449829	eDNA	jgHCO2198 - jgLCO1490 & LoboF1 - LoboR1	658	QIIME2 - Dada2	13,376	11	88.0	0.02	0.003
				PEMA	39,454	25	75.0	0.1	0.4
ERS6463899–ERS6463901				JAMP dada2 PEAR vsearch DnoisE	1,304	35	65.0	-	0.2
ERS6463906–ERS6463911 ERS6463913–ERS6463918 ERS6463920–ERS6463922	eDNA	mlCOIintF - jgHCO2198	313	PEMA	11,545	46	50.0	1	3
ERS6463744–ERS6463761				JAMP dada2 PEAR vsearch DnoisE	663	40	60.0	-	0.6
				PEMA	5,879	49	47.0	1.0	2.0
ERR3460466	bulk	mlCOIintF - jgHCO2198	313	PEMA (d = 2)	193	99	1	-	-
ERR3460467	bulk				74	97	0.0	-	3
ERR3460470	eDNA				184	71	28.0	0	1
ERS6488992	eDNA	fwhF2 - EPTDr2	142	PEMA	416	85	7	3	5
ERS6488993					315	99.2	0.4	0.4	-
ERS6488994					823	90	4	2	4
ERS6488995	eDNA	BF3 - BR2	458	PEMA	1,940	64	34.0	2	0.3

TABLE 2.7: DARN outcome over the samples or set of samples. Assignment fractions of the sequences per domain per sample in the DARN results over the samples.

[a]The *d* parameter equals 10 except mentioned otherwise

2.2.6 Discussion

By making use of a COI - oriented reference phylogenetic tree built from 1,593 consensus sequences, to phylogenetically place sequences from COI metabarcoding samples onto it, the surmise for including bacteria, algae, fungi etc. [86, 89] was verified. Our results demonstrate that standard metabarcoding approaches based on the COI gene region of the mitochondrial genome will not only amplify eukaryotes, but also a large proportion of non-target prokaryotic organisms, such as bacteria and archaea. Clearly, dark matter, and especially bacteria, make up a significant proportion of sequences generated in COI based eDNA metabarcoding datasets. The large proportion of prokaryotes observed in the present study is corroborated by the findings of [86]. Furthermore, dark matter seems to be particularly common in eDNA as compared to bulk samples [84]. However, it should be mentioned that the high number of prokaryotic sequences in COI metabarcoding data is also reflecting known issues with contamination [105, 106, 107], incorrectly labeled reference sequences [108] and holobionts [109, 110] in eukaryotic genomes.

As publicly available bacterial COI sequences are far too few to represent the bacterial and archaeal diversity, their reliable taxonomic identification is not currently possible. This way, bacterial, i.e. non-target, sequences that were amplified during the library preparation have at least the possibility of a taxonomy assignment. Our implementations using DARN indicate that it is essential both for global reference databases (e.g., BOLD, Midori etc) and custom reference databases which are commonly used, to also include non-eukaryotic sequences.

While our approach specifically addressed the COI gene, DARN can be adapted to analyse any locus fragment. For instance, metabarcoding of environmental samples for the 12S rRNA mitochondrial region is often employed to assess fish biodiversity [111, 77] and the approach presented here could be adjusted to allow further analyses of the 12S rRNA data. In addition, our approach can be used to identify non-target eukaryotes when the target is bacterial taxa [112].

The approaches implemented in DARN can benefit both bulk and eDNA metabarcoding studies, by allowing quality control and further investigation of the unassigned OTUs/ASVs. The approach is also adaptable to other markers than COI. Moreover, the approach presented here allows researchers to better understand the known unknowns and shed light on the dark matter of their metabarcoding sequence data.

Chapter 3

Studying the microbiome as a whole: the way forward

Publication relative to this chapter: ongoing work, to be submitted before phd defense, probably not accepted by then though.

3.1 Microbial interactions inference in communities of a hypersaline swamp elucidate mechanisms governing taxonomic & functional profiles

3.1.1 Introduction

3.1.2 Contribution

3.1.3 Methods

darn and PEMA will be used at this point, among other software
PREGO and dingo will be used to this end

3.1.4 Results

3.1.5 Discussion

Chapter 4

Software development to build a knowledge-base at the systems biology level

4.1 PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types[1]

Citation:
Zafeiropoulos H, Paragkamian S, Ninidakis S, Pavlopoulos GA, Jensen LJ, Pafilis E. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. Microorganisms. 2022; 10(2):293.
DOI: 10.3390/microorganisms10020293

4.1.1 Abstract

To elucidate ecosystem functioning, it is fundamental to recognize *what* processes occur in which environments (*where*) and which microorganisms carry them out (*who*). Here, we present PREGO, a one-stop-shop knowledge base providing such associations. PREGO combines text mining and data integration techniques to mine such what-where-who associations from data and metadata scattered in the scientific literature and in public omics repositories. Microorganisms, biological processes, and environment types are identified and mapped to ontology terms from established community resources. Analyses of co-mentions in text and co-occurrences in metagenomics data/metadata are performed to extract associations and a level of confidence is assigned to each of them thanks to a scoring scheme. The PREGO knowledge base contains associations for 364,508 microbial taxa, 1090 environmental types, 15,091 biological processes, and 7,971 molecular functions with a total of almost 58 million associations. These associations are available through a

[1] For author contributions and supplementary material please refer to the relevant sections. Modified version of the published review.

web interface (https://prego.hcmr.gr), an Application Programming Interface (API), and bulk download. By exploring environments and/or processes associated with each other or with microbes, PREGO aims to assist researchers in design and interpretation of experiments and their results. To demonstrate PREGO's capabilities, a thorough presentation of its web interface is given along with a meta-analysis of experimental results from a lagoon-sediment study of sulfur-cycle related microbes.

4.1.2 Introduction

Microbes are omnipresent and impact global ecosystem functions [4] through their abundance [5], versatility [113], and interactions [7]. These facts have inspired microbiologists from diverse scientific fields to study their genotype and phenotype [114], their metabolism [115], and their interactions with the environment [116]. All this work has resulted in a wealth of knowledge available in the forms of literature and experimental data. Literature contains vast amounts of information in the free text form that overwhelms researchers. Advanced text mining methods [117] have been developed to assist this issue. Experimental data and their metadata require mining [118] as well for their integration, mostly through metagenomic mining from online repositories. Hence, the combination of this knowledge about microbial life (who), their metabolic functions (what), and the environment they influence (where) is an important step to study ecosystem function [119].

High Throughput Sequencing (HTS) has turned the page on microbial ecology studies [120]. Over the past 20 years, both the taxonomic and the functional profiles of microbial communities from both local and large-scale regions (e.g., Tara Oceans [121], Earth Microbiome [122]) are being accumulated at a higher and higher rate. Extreme environments, i.e., areas with high salinity, low pH, etc., are being studied, providing us with unprecedented insight [123]. Both amplicon and shotgun metagenomics studies have played a crucial part in this development. Latest technological breakthroughs, such as Metagenome-Assembled Genomes (MAGs) and Single Amplified Genomes (SAGs), are enhancing the assessment of the taxonomic and functional repertoire of microbiomes even further. However, the mass use of these technologies and their consequent data have led to a number of needs and challenges, with metadata curation being among the most crucial ones.

Standards-promoting communities, like Genomic Standards Consortium (GSC)[2], their efforts, like Minimum Information about any (x) Sequence (MIxS) [124], and projects endorsing those, like National Microbiome Data Collaborative (NMDC) [125, 126], offer guidelines and best-practices to assist the annotation of microbial ecology samples. Controlled vocabularies and ontologies contribute to these efforts as they describe each subject area with formal terms [127]. Environment types, for example, are described by the Environment Ontology (ENVO) [128]. Other key biological aspects that have been captured include molecular functions (Gene Ontology Molecular Function (GOmf) [129, 130], Enzyme Commission nomenclature [131], etc.), and the pathways carrying out different biological processes (GO Biological Process (GObp), MetaCyc [132], etc.).

[2] https://gensc.org/

These knowledge structures, along with taxonomic centralized resources like the National Center for Biotechnology Information (NCBI) Taxonomy [133] and LPSN (List of Prokaryotic names with Standing in Nomenclature) [134], provide the means for a standardized representation of, for example, environments, process-oriented terms, and microbial taxa, respectively. Global-scale public resources (like MGnify [135], JGI/IMG [136], MG-RAST [137]) combine some of the aforementioned resources to support the collection, analysis, and distribution of multiple types of HTS data (e.g., amplicon, metagenomics, metatranscriptomics, etc.).

Besides the data and the analyses *per se*, the related scientific literature stores valuable information in billions of text lines. PubMed [133] and PubMed Central (PMC) [138] are gateways to relationships among microbes (*who*), the environments they live in (*where*) and their associated processes and functions (*what*) hidden in text [139]. Text mining (on both literature and metadata) can serve the extraction of these relationships. Named Entity Recognition (NER) can, for example, locate organism names [140], ENVO and GO terms [141] mentioned in text and map them to their corresponding identifiers. Association statistics, like co-mention analysis, can subsequently suggest ranked association among such entities [142, 143]. The new era of omics has been interwoven with data integration [144] by bringing together scattered and fragmented pieces of information.

The time is ripe for tools that integrate all this knowledge and henceforth assist researchers to tackle major challenges like climate change [8], sustainability [145], and synthetic ecology [146]. Many resources have emerged in this realm [147], each one serving a specific purpose, such as BacDive [148]. BacDive is a large-scale curated database with prokaryotic information about phenotypic, morphological, and metabolic information. Other resources like Microbe Directory [149], Web of Microbes (WoM) [150], and Microbial Interaction Network Database (MIND[3]) focus on microbial environmental conditions, metabolite interactions with microbes and microbe-microbe interactions, respectively. In addition, taking advantage of aforementioned resources, novel pipelines, e.g., [151], are emerging with the aim to explore the network associations of who (microbial taxa) is performing what (microbial processes) and where (environments) directly using graph theory [152]. These analyses and resources are important because microbiologists can enrich their data to explore hypotheses but also to identify potential gaps in knowledge regarding these associations [153].

Here, we present PREGO, a hypothesis generation web resource that is designed to be useful to microbiologists—in particular microbial ecologists and environmental microbiologists. Its specific aims include: (a) the gathering of source data, metadata, and literature followed by the extraction of microorganism, process, environment associations contained therein, (b) making such a mined knowledge base available to life sciences researchers via an easy to use and explore web portal. As such, PREGO can be useful also to system microbiologists and large-scale data analysts through bulk download and programming access. We document the principles, analysis methodology, and contents behind PREGO. Last but not least, we demonstrate PREGO's capabilities for researcher-support related to the above through a case study involving sulfate-reducing microorganisms.

[3] http://www.microbialnet.org/mind_home.html

4.1.3 Contribution

4.1.4 Methods & Implementation

PREGO is a resource designed to assist molecular ecologists in acquiring a single point overview of what-where-who process–environment–organism associations. The system is comprised of two main parts: (a) a server that periodically harvests data and extracts process-environment-organism associations from the scientific literature, environmental samples, and genome annotation sequences (Figure 4.1, step 1 to 5) and (b) a web-based interface as well as an Application Programming Interface (API) that provides users and programmers with a friendly way to extract and navigate across the process–environment–organism associations (Figure 4.1, step 6).

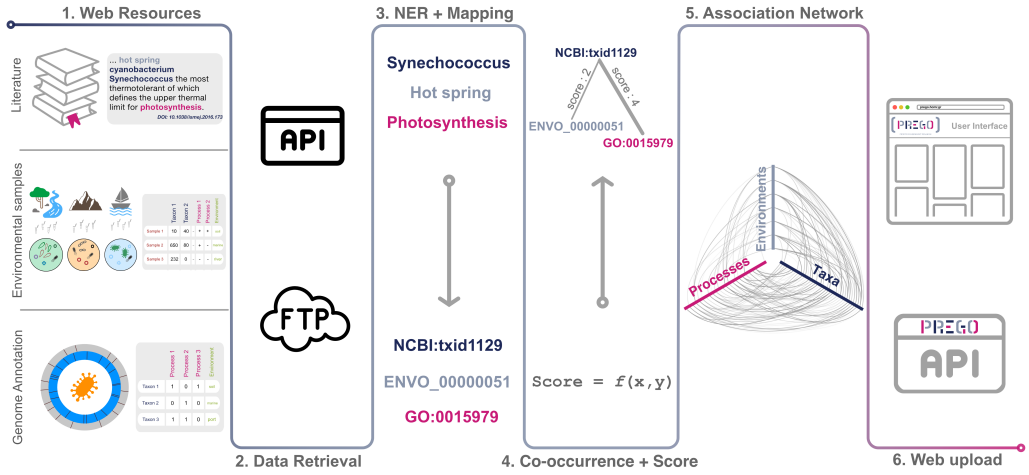

FIGURE 4.1: PREGO analysis methodology: PREGO periodically retrieves three distinct types of data from open access resources. Scientific text, environmental sample data, and genomic annotations are handled with respective methodologies in order to standardize their entities. Named Entity Recognition and Comention/Co-occurrence analysis is the common framework in order to build a weighted association network with nodes being the entity identifiers. Lastly, all these associations are available through a Web interface and an API. All these steps have been implemented in an autonomous way with regular cycles of updates (see Appendix A.2). Icons used from the Noun Project released under CC BY: Books by Shakeel Ch., Bacteria by Maxim Kulikov, ftp by DinosoftLab, Mountain by Diane, Ship on Sea by farra nugraha, River by Chanut is Industries.

Entity Types, Channels, and Associations

PREGO supports three entity types: *Process*, *Environment*, and *Organism*. For interoperability and consistency, an ontology or taxonomy is adopted for each type of entity. Processes are represented as Gene Ontology (GO) terms and are grouped either as Biological processes (GObp) or as Molecular functions (GOmf). In addition, Environments are represented by terms from the Environmental Ontology. Organisms are represented by

the microbial NCBI Taxonomy Ids (Bacteria, Archaea, and unicellular eukaryotes). For the unicellular eukaryotes, a custom list was populated with the unicellular eukaryotic taxa using a curated list. PREGO's contents are mainly divided into three distinct channels of information based on data origin and format (Figure 4.1, step 1). The *Literature* channel exploits scientific publications, i.e., abstracts and full text open access scientific publications (Table 4.1 and Section 4.1.4). Through the *Annotated Genomes and Isolates* channel, PREGO retrieves genome annotations and their accompanying metadata (Table 4.1 and Section 4.1.4). Finally, the *Environmental Samples* channel supports the integration of metagenomic analyses from both amplicon and shotgun studies. These include taxonomic and functional profiles along with their corresponding metadata (Table 4.1, more details in Section 4.1.4).

Source	# items	Data type	Metadata	License
MEDLINE and PubMed	33 million	abstracts (text)	no	NLM Copyright
PubMed Central OA Subset	2.7 million	full article (text)	no	CC for Commercial, non-commercial
JGI IMG	9,644	Isolates Annotated genomes	yes	JGI Data Policy
Struo	21,276	Annotated genomes	no	MIT, CC BY-SA 4.0
BioProject	18,752	Annotated genomes with abstracts (text)	yes	INSDC policy
MG-RAST	16,096	markergene samples	yes	CC0
	7,965	metagenomic samples	yes	CC0
MGnify	10,500	markergene samples	yes	CC-BY, CC0

TABLE 4.1: Source databases that are integrated in PREGO and the number of items retrieved. The Open Access subset of PubMed Central has a Creative Commons license available for commercial and noncommercial use. JGI has its own license, the same applies for BioProject, MEDLINE®, and PubMed® as well.

In cases in which the retrieved data and metadata are in text form, they are standardized to the aforementioned identifiers and taxonomies using Named Entity Recognition (NER) tools, namely the EXTRACT tagger [141, 154]. In cases where data contain KEGG Orthology terms and/or Uniref identifiers, they are mapped to the respective GOmf using the mapping files available from the UniProt (see Appendix A.1). Associations are extracted after the mapping and standardization of the entities from each resource (Figure 4.1, step 3). The association extraction pipeline is distinct for each channel and resource because of differences in the data type origin (see `prego_gathering_data` in the Availability of Supporting Source Codes section). By the means of navigation, the large number of associations returned to the user require a type of sorting; ideally, one that ranks the most trustworthy associations at the top. For those reasons, each channel of PREGO has a dedicated scoring scheme bounded within the (0,5] space for consistency. In Appendix A.1 , the scoring scheme of each channel is elaborated.

Text Mining of Scientific Literature

PREGO implements a text mining methodology to extract associations of the aforementioned entities from literature. The origin of text mining is a corpus that comprises scientific abstracts and full text articles from MEDLINE® and PubMed® and PubMed Central® Open Access Subset (PMC OA Subset) [155], respectively. The building and periodic update of the corpus is possible through the NCBI File Transfer Protocol (FTP) services. PREGO also has a dedicated text-mining dictionary (see Availability of Supporting Source Codes section) that contains the entities ids, names, synonyms, and neglected words (stop words). PREGO dictionary incorporates the ORGANISMS [140] and ENVIRONMENTS [156] evaluated dictionaries as well as the experimental dictionaries of Gene Ontology Biological Process and Molecular Function. Text mining is subsequently performed on the corpus using the dictionary through the EXTRACT tagger [141, 154]. The tagger recognizes the entities of the dictionary in each abstract and full text article and assigns their co-mentions with a score. The score is sensitive to the text structural level of co-mention; higher to lower scoring when co-mention appears in the same sentence, then, in the same paragraph, and lastly in the same article. All these are integrated and normalized to a single score for each association, as implemented in STRING 9.1 [143] (see Appendix A.3 for more details). In addition, the tagger extracts each mention in every article to provide the origin of each association it extracts.

Annotated Genomes and Isolates

Annotated genomes and isolates comprise the most trustworthy data in PREGO's knowledge base because they refer to a single species/strain and also have manually curated metadata. Among other data types, JGI-IMG [136, 157] includes millions of genes from isolated genomes (isolates), SAGs and MAGs. Such annotations, along with their corresponding metadata, were collected using web-parsing technologies. Their metadata, describing their related environment/ecosystem, were tagged using the EXTRACT tagger to infer organisms—environments associations. The annotated KEGG terms were mapped to GOmf terms (see Appendix A). The GOmf terms were then used to extract organisms—processes associations.

The Struo pipeline [158] and its outcome when using the Genome Taxonomy DataBase (GTDB) (v.03-RS86) [159] was exploited to enrich organisms—processes associations. A set of 21,276 representative genomes, accompanied by UniRef50 annotations, was retrieved using the provided FTP server. The annotations were then mapped to GOmf terms (see Appendix A.1). Related GTDB genomes were mapped to their corresponding NCBI taxa (see Appendix A.1). All associations extracted from these resources were assigned arbitrarily a confidence level of four out of five. This score choice reflects the high-quality of these data and metadata.

In addition, BioProject data were integrated to PREGO using the NCBI FTP/e-utils services [155]. The BioProject ids that were integrated are the ones that have been assigned a PubMed abstract, a unicellular taxon, and Genome sequencing as data type. Then, using the text mining pipeline, associations were extracted connecting the assigned taxon with the rest of the entities that appear in the abstracts. This method resulted in associations

that were assigned a confidence level of three (out of five) because of the combined method of curated data with text mining.

Environmental Samples

MGnify [135] and MG-RAST [137] repositories provide a great number of public metagenomic records. In the PREGO framework, both amplicon and shotgun metagenomic analyses are retrieved periodically along with their corresponding metadata. Data retrieval from these resources is possible from their Application Programming Interfaces (APIs). Marker gene analyses are retrieved and by measuring the co-occurrence of taxa present in the various environmental types (e.g., biomes, materials, features, etc.) organisms—environments associations are extracted. These associations emerge when a taxon is reported together with a certain environmental type, being mentioned in the metadata of a sample (metadata based co-occurrence). Similarly, analyses of metagenomic samples along with their corresponding metadata and annotations are also retrieved and organisms—environments, organisms—processes and processes—environments are extracted. The processes—environments associations are possible through co-occurrence of the functional annotations of metagenomes with the environmental metadata of the samples.

In all cases, the EXTRACT tagger is used on the microorganism names and the corresponding metadata of each sample to identify their identifiers (NCBI ids, ENVO terms, GOmf, GObp). All associations in this channel are scored based on the number of samples the entity of interest co-occurs with specific sample metadata (e.g., environmental type) or annotations (functional annotations or taxonomic annotations). The same scoring scheme was implemented across the channel resources (see Appendix A.3 for more details), which ranks these associations with a value in the (0,5] space.

Sequence Search

In the case of organisms, PREGO enables sequence-based queries, meaning a sequence (amplicon) can be used as an entry point like it was a taxon name. To this end, a custom database was built using a set of reference custom databases for four commonly used marker genes. For 16S and 18S rRNA, the SILVA database (v.138) [37] and the PR2 database (version_4.14.0) [160, 161] were used. Cytochrome c oxidase I (COI) [162] is another commonly used marker gene; for this reason, Midori 2 (version GB243) [163] was integrated in PREGO's custom database. Finally, for the Internal transcribed spacer (ITS), common in studies focusing on Fungi, the Unite (version 8.3, accessed 10.05.2021) [52] database was added.

Back-End Server and Front-End Implementation

PREGO is a multi-tier web-based application. It is hosted on a 64 GB RAM DELL R540, 20 core, Debian server. Custom API clients (written in Python) are responsible for retrieving the data and metadata from each source (Figure 4.1, step 2). These clients as well as the subsequent methodology (Figure 4.1, step 3 to 6) are updated in regular cycles using custom daemons (see Appendix A.2, Figure A.1). The *mamba/blackmamba* web framework underlies communication to the Postgres association-holding database and

the client-side communication. HTML 5, Ajax, JQuery, and custom Javascript enhance the user web experience. PREGO supports widely used browsers (e.g., Chrome, Firefox, Safari, Edge) in various operating systems, such as Windows 10, Linux (Ubuntu 18), and MacOS (10.12, 11).

4.1.5 Results & Validation

The PREGO Web Resource

Users can access the PREGO contents through its web User Interface (UI) (Figures 4.2 and 4.3), its Application Programming Interface (API) (Figure 4.4), or bulk download of all associations (Appendix D). The User Interface comes with two search fields: a plain text search and a sequence search (Figure 4.2a). The latter is used when the user wants to search for a taxon sequence (see Section 4.1.4 for supported sequence databases). The plain text search supports three types of entry points; the user can search for a taxon name, e.g., *Methanosarcina mazei*, an environmental type, e.g., lagoon, or a biological process e.g., methanogenesis. In all entry points, PREGO returns an overview page consisting of tabs with associations of the entity of interest with the entities of the two other types (Figure 4.2b–d) as well as Documents and Downloads tabs (Figure 4.2e,f).

Regarding the association tabs, when a taxon is used as a query, PREGO returns an overview page consisting of tabs for environments, biological processes, and molecular functions. When an environmental type is used as input, PREGO returns the organisms that have been found to be related to it, as well as the Biological Processes observed in the given environment. Lastly, if a biological process is under study, PREGO returns a tab with the organisms along with another tab with the Environments related to the process. Notably, only the associations with scores higher than 0.5 are presented in the web platform and are sorted in descending order based on their score. The score is visualized with a five-star system (see Appendix A.3 for the scoring scheme). Every association tab contains three tables with associations derived from the PREGO channels (see Section 4.1.4) along with their supported evidence. The user can both search and scroll through these tables, which makes knowledge extraction easier in cases where, for example, Isolate data contain hundreds of associations. In the *Literature* channel, each association is supported by the scientific articles with text-mining identified co-mentions. When a user clicks on an association, a popup window appears. This window displays abstracts or excerpts of full text with the associated entities highlighted (Figure 4.3a). Additionally, the Environmental Samples and Genome annotations and Isolates channels support evidence for each association by providing links to more detailed information. In the former channel, when the users click on an association, they are redirected to pertinent sample pages of MGnify (Figure 4.3b). Similarly, the latter redirects users to JGI and NCBI genomes when the associations originated from JGI—IMG and Struo, respectively (Figure 4.3c).

The *Documents* tab includes a list of scientific publications where the queried entity is mentioned. Through the *Downloads* tab, users are able to get all of the PREGO associations found for their query, per entity type (e.g., all the environments found related to an organism) and per channel (e.g., all the Environments found related to an organism

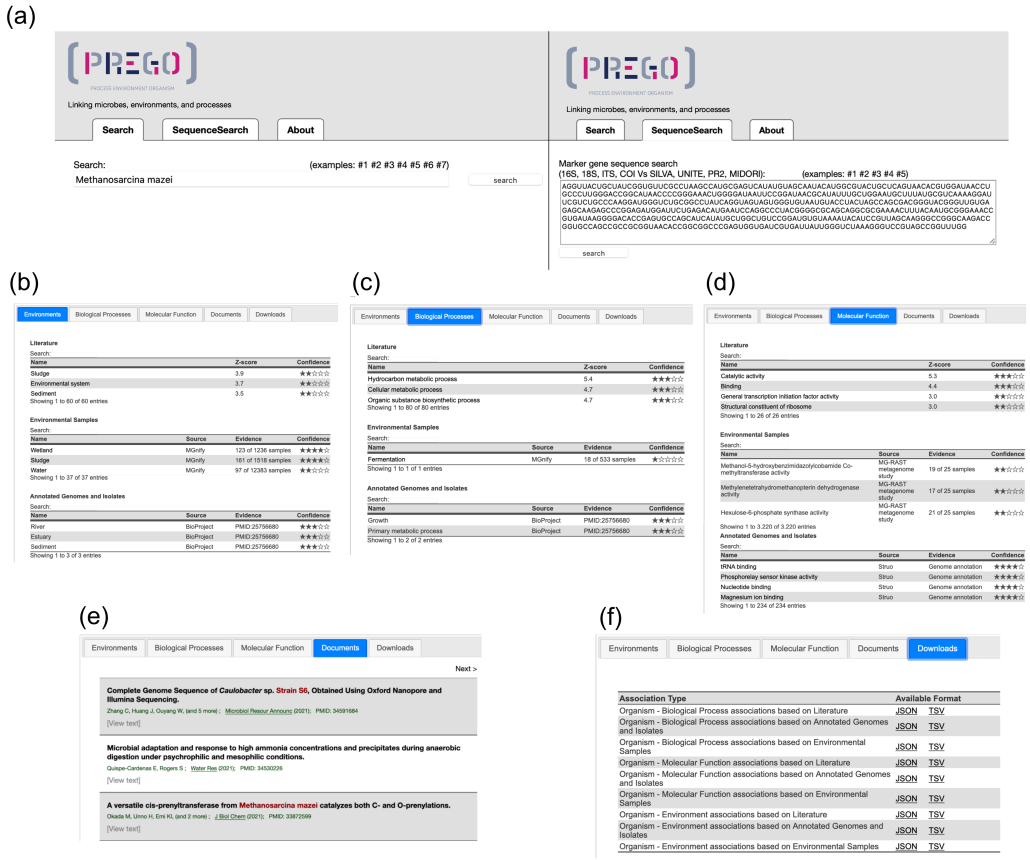

FIGURE 4.2: PREGO web user interface. (a) There are two search fields, plain text and taxa sequences. (b-d) three associations tabs each one presenting associations of the querred entity with the respective entities, Environments (b), Biological Process (c) and Molecular Function (d). Three channels of information are distinguishing the associations based on the original data. (e) Documents tab presents the scientific articles that mention the queried entity highlighted with color. (f) Downloads tab provides the associations of each channel (when available) to be downloaded in JSON and TSV format.

through the *Literature* channel). This data retrieval functionality is also available via the PREGO API (syntax described in Figure 4.4). Finally, all PREGO associations are available for bulk download from each channel (see Table A.2).

PREGO in Action

To demonstrate PREGO's potential, we present four different ways that PREGO can assist molecular ecologists. The demo focuses on the sulfate-reducing microorganisms (SRMs) as well as the processes and environments that relate to sulfate reduction. Through this demo, we highlight how the different channels may provide complementary insights regarding different taxonomic levels and different association types.

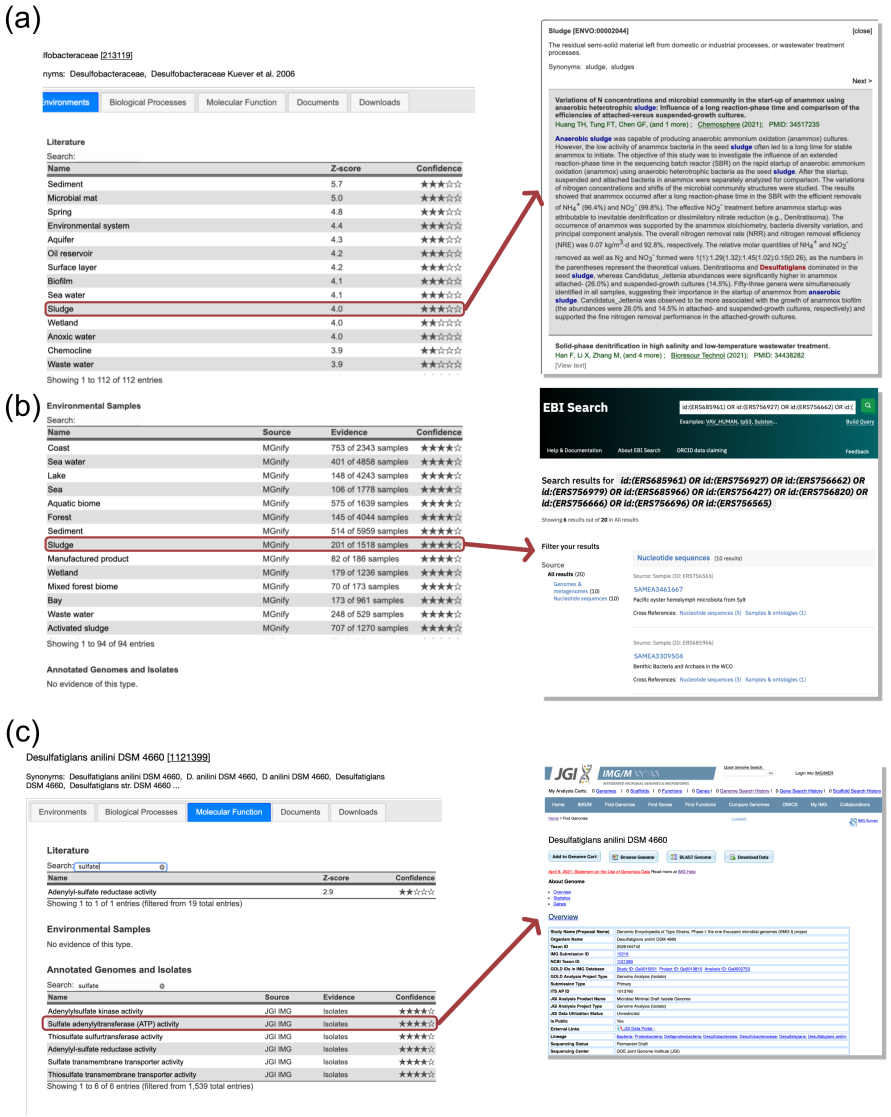

FIGURE 4.3: Each association is supported by original data. (a) Literature channel has a pop-up functionality that displays the scientific articles that each specific association occurs with highlighted color. (b) Environmental Samples channel redirects to the samples that support the specific association (currently only is supported MGnify). (c) Annotated Genomes channel similarly redirects to the isolates ids that each association is based on (both Struo and JGI IMG are supported).

Which Environments Are Related to a Taxon?

Based on Pavloudi et al. (2017) [164], several bacterial and archaeal SRM were found in lagoonal sediments, after amplifying and sequencing the dissimilatory sulfite reductase β-subunit (dsrB). Using PREGO for the case of Desulfobacteraceae, the family in which the majority of the observed OTUs of the study belonged to, several environmental types

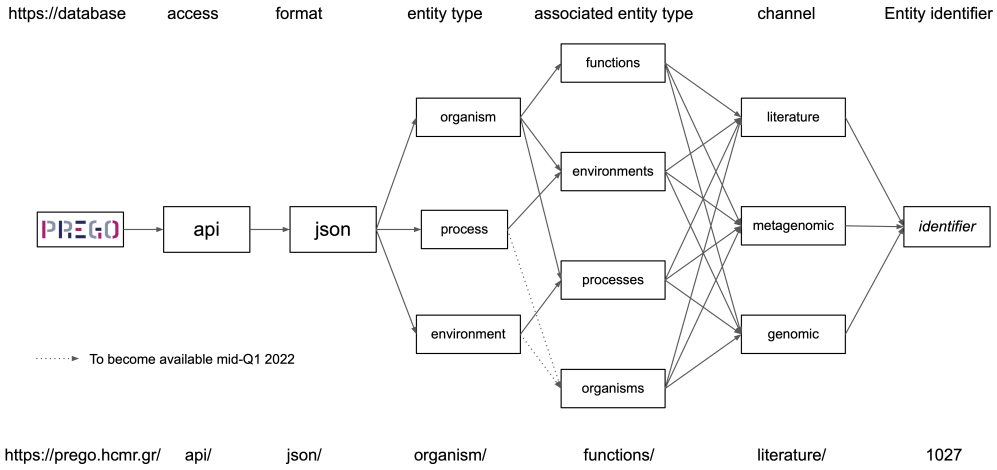

FIGURE 4.4: The PREGO API schema.

similar to lagoons were retrieved from both the *Literature* and the *Environmental samples* channels (Figure 4.3a,b). Moreover, most of them had a high z-score, such as *"sediment"*, *"sludge"*, and *"activated sludge"*. Several dissimilar environmental types were associated with Desulfobacteraceae, e.g., *"oil reservoir"* indicating them as potential environments where sulfate reduction takes place. However, the presence of taxa within that family in different environments, from *"sea water"* to *"forest"* and *"Wastewater treatment plant"*, may suggest that this family has ubiquitous representatives in diverse conditions.

Searching for *Desulfatiglans anilini* (example1[4], accessed on 24 December 2021) at the species level, the most abundant species in Pavloudi et al. (2017) and, for *Desulfatiglans anilini* DSM 4660 strain (example 2[5], accessed on 24 December 2021), PREGO provides associations with the *"Anaerobic sediment"*, *"Marine oxygen minimum zone"*, and *"Anaerobic digester sludge"* terms. These associations further corroborate the relationship between the species and sulfate reduction. More specifically, the *"sulfur spring"* ENVO term was retrieved from the Environmental samples channel as well.

Which Biological Processes and Molecular Functions Are Related to a Taxon?

According to Pavloudi et al. (2017), *Desulfatiglans anilini* plays an important role in sulfate reduction. The Biological Processes provided by PREGO's Literature channel are the GO terms *"Sulfate reduction"*, *"Sulfide oxidation"*, and *"Sulfide ion homeostasis"*, which support this claim. In addition, the *"Denitrification pathway"* term was also retrieved. This is rather interesting as it is in line with what Pavloudi et al. (2017) discussed about the SRMs and their ability to use various electron acceptors, e.g., nitrate and nitrite.

[4] https://prego.hcmr.gr/example1
[5] https://prego.hcmr.gr/example2

Furthermore, PREGO's Molecular Function tab provides more insight on this example. Several GO terms related to sulfate reduction (e.g., terms related to *"sulfite reductase"*) were associated with DSM 4660 strain and Desulfatiglans anilini species in multiple channels. Interestingly, in the case of the strain query, the Annotated Genomes channel returned many GO terms related to the nitrogen fixation, e.g., *"nitric oxide dioxygenase activity"*.

Which Taxa Are Related to a Biological Process?

PREGO can be also used to report organisms that relate to a certain biological process. Searching for *"dissimilatory sulfate reduction"* associations with taxa (example 3[6], accessed on 24 December 2021) resulted in several taxa that were mentioned in the Pavloudi et al. (2017) study. For example, taxa such as Thermodesulfobacteria and Thermodesulfovibrio were found among the entries with the highest score (e.g.,) based on the Literature channel. The other two channels did not contain any associations. Using the *"Sulfate assimilation"* (example 4[7], accessed on 24 December 2021) as the biological process input, PREGO results showed several genera that were missing from PREGO results concerning the *"dissimilatory sulfate reduction"*. Hence, manual search of GObp terms that describe the actual biological process of interest is more insightful.

Are There Any Associations between Environments and Biological Processes?

Are there other environmental types, except the lagoonal sediments, in which sulfate assimilation occurs? In that question, and in *"dissimilatory sulfate reduction"* (example 3) in particular, PREGO assigns the highest score to "sediment" while, among others, *"anoxic water"*, *"oil reservoir"*, *"mud volcano"*, and *"basalt"* are potentially associated with environments related to sulfate reduction.

Inversely, PREGO is insightful about occurring processes in a specific environmental type. For example, searching for the biological processes that take place in *"basalt"* (example 5[8], accessed on 24 December 2021), processes like *"Nitrogen fixation"* and *"Reactive nitrogen species metabolic process"* stand out. However, sulfate reduction is not among the associations. However, when asking for *"Mafic lava"* (example 6[9], accessed on 24 December 2021), both the *"nitrogen fixation"* and *"Sulfur compound metabolic process"* terms are returned. This highlights the suggestions of Pavloudi et al. (2017), regarding the potential use of various electron acceptors from the different strains present in different environmental types.

PREGO Contents

PREGO contains the literature, environmental samples, and genome annotations of the resources shown in Table 4.1. The extracted contents of these resources have resulted to a knowledge base with 364 K distinct taxonomic groups (out of a pool of 620K Bacteria,

[6] https://prego.hcmr.gr/example3
[7] https://prego.hcmr.gr/example4
[8] https://prego.hcmr.gr/example5
[9] https://prego.hcmr.gr/example6

Archaea, and microbial eukaryotes, based on NCBI Taxonomy) from which 258K are at the species level (Table 4.2). These taxa are associated with 1 K Environment Ontology terms, 15 K GObp terms, and with 7.9 K GOmf terms. Combining the above, PREGO maintains a knowledge base of entities and associations between them that form a multipartite network with entities as nodes and scored associations between them as weighted links.

Channel	Source	Taxonomy		Environments	Biological Processes	Molecular Functions
Literature	MEDLINE PubMed - PMC OA	Strains	8,929	1,077	15,079	7,318
		Species	240,377			
		Total	342,506			
Environmental samples	MG-RAST amplicon	Strains	1,392	162	-	-
		Species	4,324			
		Total	5,859			
	MG-RAST metagenome	Strains	2,522	258	-	3,839
		Species	4,406			
		Total	7,157			
	MGnify amplicon	Strains	2	216	11	-
		Species	1,471			
		Total	2,955			
Annotated Genomes & Isolates	JGI IMGisolates	Strains	2,398	241	-	3,670
		Species	11,203			
		Total	13,849			
	STRUO	Strains	6	-	-	2,789
		Species	19,289			
		Total	19,325			
	BioProject	Strains	5,754	309	626	-
		Species	3,373			
		Total	9,393			
Total	All	Strains	12,840	1,090	15,091	7,971

TABLE 4.2: The entities of PREGO after the NER and mapping of every source. Counts of distinct entities of Taxa, Environments (ENVO terms), Biological Processes (Gene Ontology Biological process) and Molecular Function (Gene Ontology Molecular Function).

As shown in Figure 4.5, in its current version (December 2021), PREGO knowledge base covers 157 bacterial phyla (107 are Candidatus), 23 phyla from archaea (18 are Candidatus), and 22 unicellular eukaryotic phyla described in the NCBI Taxonomy database. The number of bacterial taxa present among the associations of each phylum ranges from the order of 10s, as in the case of *Candidatus Coatesbacteria*, to hundreds of thousands, e.g., Actinobacteriae. The number of environmental types, found among the PREGO associations for each phylum, ranges from just a few to up to 1000. Similarly, the number

of biological processes that have been related to the various phyla may range from less than a dozen, e.g., Yanofskybacteria to up to several thousands, e.g., Bacteroidetes. On the contrary, the number of molecular functions found to be related to taxa of each phylum is rather constant in all three domains.

4.1.6 Discussion

PREGO Contents

On its current version and according to the NCBI Taxonomy that it is based on, PREGO manages to cover a great range of microbial taxa, as most (if not all phyla) are present in the knowledge base (Figure 4.5). The different number of organisms' entities per phylum highlights the diverse number of the members of the various phyla. On the contrary, the similar number of molecular functions in all cases indicates the robustness of the main metabolic processes required for life. With respect to biological processes, their number per phylum varies to some extent, especially for the case of Bacteria and Archaea. That could be observed as, in many cases, phyla that have been recently described using molecular techniques have not been studied extensively yet, e.g., Candidatus Delongbacteria. As expected, the number of environmental types that have been associated with members of each phylum varies, as a phylum may be universally present, while others could be strongly niche-specific (e.g., Hydrothermarchaeota).

Because of its three different channels, PREGO manages to extract associations both in the species and higher taxonomic levels. The Isolates channel supports explicit associations at the species level (Table 4.3 and Figure S3). Interestingly, the number of such genomes seems to have reached a plateau for now, as PREGO-like platforms include the same order of magnitude. The *Literature* channel, on the other hand, promotes the extraction of associations at higher taxonomic levels (Table 4.3 and Figure S1). This also applies to environment—organisms associations derived from the Environmental Samples channel (Table 4.3 and Figure S2). Associations regarding biological processes, though, are strongly enhanced by the Literature channel and the massive increase of literature.

Additionally, the text mining methodology of the Literature channel has retrieved most of the entities present in PREGO knowledge base (Table 4.2). A significant contribution to the taxa with associations is due to the PMC OA processing by the text mining pipeline of the Literature channel. This is in-line with reports in other applications of text mining when using full text articles [165]. However, the resulting associations are suggestive because of the text mining nature, and therefore subject for further review by the users.

Related Tools' Functionality and Content

There is an emerging niche for tools similar to PREGO to bring forward microbe associations and metadata. Table 4.4 summarizes the common and different features of BacDive, WoM, NMDC data portal, and PREGO. All of them commonly share the environmental associations and biological/metabolic processes with the microbes.

BacDive is a well-established platform with a focus on phenotype and cultivation information for about 100,000 prokaryotes, bacteria, and archaea. It has a high level

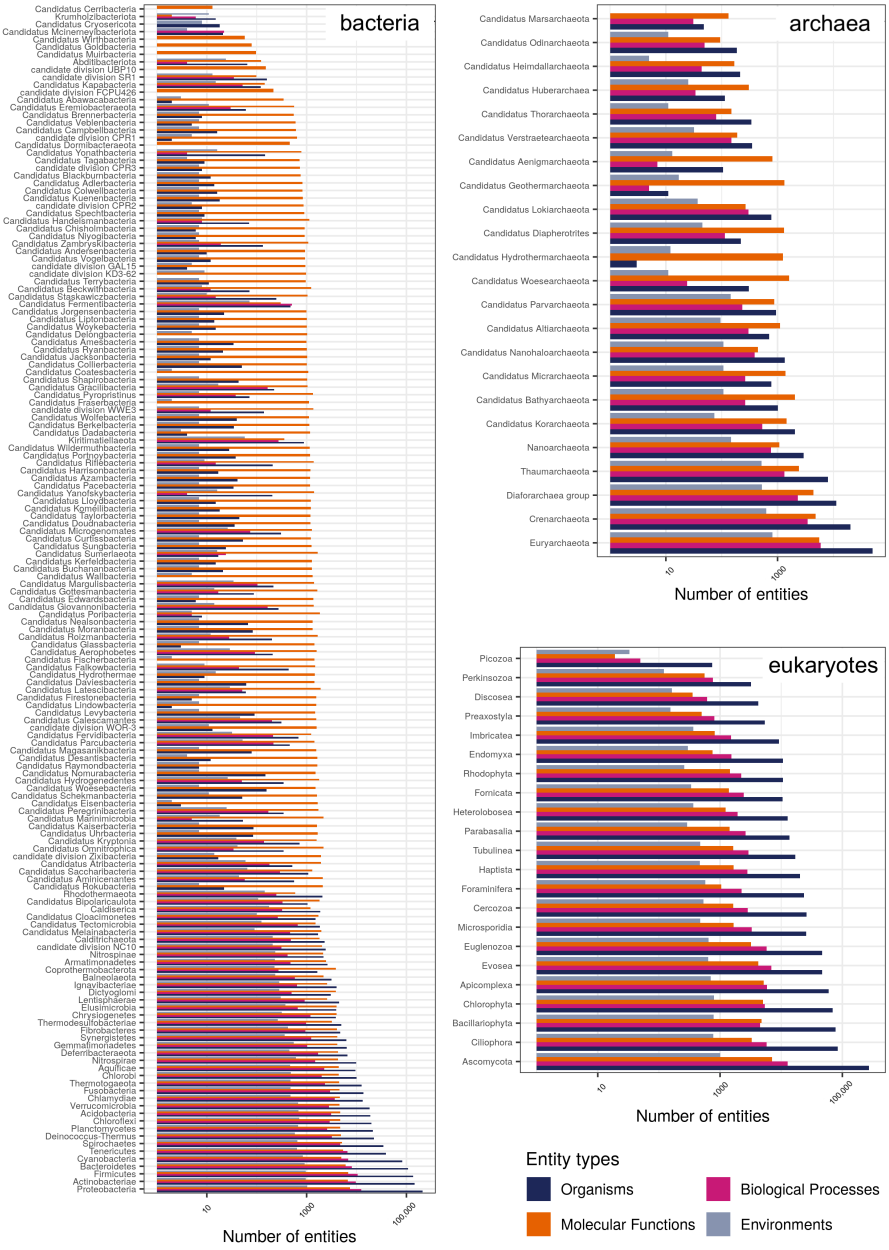

FIGURE 4.5: Summary of all the unique entities per phylum for each of the four entity types (in log10 scale) that appear in PREGO. Phyla are grouped based on their superkingdom (in log10 scale). Only phyla for which associations are available in the PREGO platform are mentioned.

Channel	Source	Environments - Processes	Environments - Functions	Taxonomy	Taxa - Environments	Taxa - Processes	Taxa - Function
Literature	MEDLINE PubMed - PMC OA	883,997	422,579	Strains	69,968	590,630	384,079
				Species	778,877	3,501,635	1,961,920
				Total	1,669,608	7,969,310	4,613,827
Environmental samples	MG-RAST amplicon	-	-	Strains	13,645		
				Species	39,007	-	-
				Total	53,439		
	MG-RAST metagenome	-	620,846	Strains	262,106		8,626,328
				Species	103,913	-	10,715,548
				Total	372,301		19,950,096
	MGnify amplicon	-	-	Strains	18	-	
				Species	30,122	351	-
				Total	111,976	2,097	
	JGI IMG isolates	-	-	Strains	8,229		3,461,693
				Species	42,141	-	13,216,559
				Total	50,888		16,821,850
Annotated Genomes and Isolates	STRUO	-	-	Strains			1,803
				Species	-	-	4,070,195
				Total			4,079,312
	BioProject	-	-	Strains	3,263	7,473	
				Species	4,187	4,294	
				Total	7,641	12,169	
Total	All	883,997	1,043,425	Strains	357,229	598,103	12,473,903
				Species	998,247	3,506,280	29,964,222
				Total	2,265,853	7,983,576	45,465,085

TABLE 4.3: The associations between entities of PREGO after co-occurrence analysis: The supported entity types of associations are Environments—Biological Processes, Environments—Molecular Functions, Taxa—Environments, Taxa—Biological Processes, Taxa—Molecular Functions.

of curation for most of its input types, like literature, internal databases, and personal collections. The NMDC data portal has published the scheme, the user interface, and a demonstrative collection of samples that will be populated later on. Standout features are the spatial visualization with coordinates and the detailed information of the samples, e.g., sequencing instruments and methodology. An alternative approach is facilitated by WoM, which aims to bind chemistry to microbes. An environment, in particular, is defined as the starting metabolite pool that is transformed by an organism. Another tool is The Microbe Directory that contains fully curated metadata for about 8000 microbes from all superkingdoms. This tool focuses on conditions of growth and on host taxa.

Complementary to these tools, PREGO contains associations of bacteria, archaea, and eukaryotes. Distinctive features are the associations of environments with processes/functions and the large-scale literature integration with text mining. Most importantly, most of the tools are complementary to each other with minimum overlap, an indication of the opportunities for further innovative synergies.

Functionality	BacDive	Web of Microbes	NMDC	PREGO
manual curation	high	high	intermediate	low
literature integration	limited	no	no	yes
environment—taxa associations	yes	yes	yes	yes
environment—process/ function associations	no	no	no	yes
process/function—taxa associations	yes	yes	yes	yes
phenotypic data	yes	no	no	no
data origin	original integration	original	original integration	integration
spatial coordinates	yes	no	yes	no
application programming interface	yes	no	yes	yes
bulk download	limited	yes	yes	yes

TABLE 4.4: Feature comparison among platforms that facilitate knowledge discovery and integration of microbial data.

PREGO Next Steps

PREGO is a user-friendly association mining and sharing platform. Its modular web-architecture grants it the flexibility for further improvements in the aforementioned aspects, namely: source datasets, user interface, entity, and association scope expansion. Regarding datasets, additional data, such as transcriptomes from MGnify and other records annotated with metadata from studies in EuroPMC, accessed on 24 December 2021) [166], could be incorporated. Similarly, the NMDC data platform standards-compliant annotated records[10] (accessed on 24 December 2021) could serve as an additional resource

[10] https://data.microbiomedata.org/

with its high-quality metadata [125, 126]. Reciprocally, if requested, pertinent literature and association summaries could be programmatically offered to interested third parties.

Furthermore, the entity types supported by the PREGO system could be expanded. For example, GOmf terms could be upgraded as a search-entry point to the system. In addition, disease and tissue describing terms, already supported by the PREGO-underlying EXTRACT system [141], could enter the PREGO ecosystem of associated entities. From a statistics perspective, the calculation of a combined association score, when an association is reported by more than one channel of information, could be another feature to add.

The user interface can be enhanced to support multiple entity and/or sequence queries, instead of single ones. Sequences can be processed by taxonomy assignment pipelines (e.g., PEMA [101]) and be converted into searching PREGO for associations. In addition, network visualization tools, like Arena3Dweb [167], could allow interactive browsing of associations through multi-layered graphs. Enrichment analyses, like those performed by OnTheFly2.0 [168] or Flame [169], can be incorporated. Omics data analysis pipelines, like MiBiOmics [170], environment associations with sequences using SeqEnv [171] and biogeochemical associations with metagenomic data with DiTing [172] could be enabled by comparing the associations pertinent to different groups of entities. The computationally intensive tasks of multiple queries, taxonomy assignments to sequences and enrichment analysis could be offered by our in-house High Performance Computing facility (https://hpc.hcmr.gr/, accessed on 24 December 2021) [98] in synergy with pertinent Research Infrastructures like ELIXIR[11] (accessed on 24 December 2021) and LifeWatch ERIC[12] (accessed on 24 December 2021).

Availability of Supporting Source Codes:

The PREGO software modules are available under BSD 2-Clause “Simplified” License. Scripts, where additional libraries have been used, are subject to their individual licenses. More information on each module can be found as listed below:

- prego_gathering_data github.com/lab42open-team/prego_gathering_data
- prego_daemons github.com/lab42open-team/prego_daemons
- prego_mappings github.com/lab42open-team/prego_mappings
- prego_statistics github.com/lab42open-team/prego_statistics

Additional software and curated lists along with their individual license are:

- tagger: https://github.com/larsjuhljensen/tagger, BSD 2-Clause "Simplified" License
- mamba: https://github.com/larsjuhljensen/mamba, BSD 2-Clause "Simplified" License

[11] https://elixir-europe.org
[12] https://www.lifewatch.eu/

- tagger dictionary: https://download.jensenlab.org/ and there in: https://download.jensenlab.org/prego_dictionary.tar.gz, CC-BY 4.0 license

Chapter 5

Software development to establish metabolic flux sampling approaches at the community level

5.1 A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks

Citation: Chalkis, A., Fisikopoulos, V., Tsigaridas E. and Zafeiropoulos H.
Geometric Algorithms for Sampling the Flux Space of Metabolic Networks.
37th International Symposium on Computational Geometry (SoCG 2021)
DOI: 10.4230/LIPIcs.SoCG.2021.21

5.1.1 Abstract

Systems Biology is a fundamental field and paradigm that introduces a new era in Biology. The crux of its functionality and usefulness relies on metabolic networks that model the reactions occurring inside an organism and provide the means to understand the underlying mechanisms that govern biological systems. Even more, metabolic networks have a broader impact that ranges from resolution of ecosystems to personalized medicine.

The analysis of metabolic networks is a computational geometry oriented field as one of the main operations they depend on is sampling uniformly points from polytopes; the latter provides a representation of the steady states of the metabolic networks. However, the polytopes that result from biological data are of very high dimension (to the order of thousands) and in most, if not all, the cases are considerably skinny. Therefore, to perform uniform random sampling efficiently in this setting, we need a novel algorithmic and computational framework specially tailored for the properties of metabolic networks.

We present a complete software framework to handle sampling in metabolic networks. Its backbone is a Multiphase Monte Carlo Sampling (MMCS) algorithm that unifies rounding and sampling in one pass, obtaining both upon termination. It exploits an improved variant of the Billiard Walk that enjoys faster arithmetic complexity per step. We demonstrate the efficiency of our approach by performing extensive experiments on vari-

ous metabolic networks. Notably, sampling on the most complicated human metabolic network accessible today, Recon3D, corresponding to a polytope of dimension 5335, took less than 30 hours. To our knowledge, that is out of reach for existing software.

5.1.2 Introduction

The field of Systems Biology

Systems Biology establishes a scientific approach and a paradigm. As a research approach, it is the qualitative and quantitative study of the systemic properties of a biological entity along with their ever evolving interactions [173, 174]. By combining experimental studies with mathematical modeling it analyzes the function and the behavior of biological systems. In this setting, we model the interactions between the components of a system to shed light on the system's *raison d'être* and to decipher its underlying mechanisms in terms of evolution, development, and physiology [175].

Initially, Systems Biology emerged as a need. New technologies in Biology accumulate vast amounts of information/data from different levels of the biological organization, i.e., genome, transcriptome, proteome, metabolome [176]. This leads to the emerging question *"what shall we do with all these pieces of information"?* The answer, if we consider Systems Biology as a paradigm, is to move away from reductionism, still the main conceptual approach in biological research, and adopt holistic approaches for interpreting how a system's properties emerge [177]. The following diagram provides a first, rough, mathematical formalization of this approach.

$$\textit{components} \rightarrow \textit{networks} \rightarrow \textit{in silico models} \rightarrow \textit{phenotype}\ [178].$$

Systems Biology expands in all the different levels of living entities, from the molecular, to the organismal and ecological level. The notion that penetrates all levels horizontally is *metabolism*; the process that modifies molecules and maintains the living state of a cell or an organism through a set of chemical reactions [179]. The reactions begin with a particular molecule which they convert into some other molecule(s), while they are catalyzed by enzymes in a key-lock relationship. We call the quantitative relationships between the components of a reaction *stoichiometry*. Linked reactions, where the product of the first acts as the substrate for the next, build up metabolic pathways. Each pathway is responsible for a certain function. We can link together the aggregation of all the pathways that take place in an organism (and their corresponding reactions) and represent them mathematically using the reactions' stoichiometry. Therefore, at the species level, metabolism is a network of its metabolic pathways and we call these representations *metabolic networks*.

From metabolism to computational geometry

The complete reconstruction of the metabolic network of an organism is a challenging, time consuming, and computationally intensive task; especially for species of high level of complexity such as *Homo sapiens*. Even though sequencing the complete genome of a species is becoming a trivial task providing us with quality insight, manual curation is still

mandatory and large groups of researchers need to spend a great amount of time to build such models [180]. However, over the last few years, automatic reconstruction approaches for building genome-scale metabolic models [181] of relatively high quality have been developed. Either way, we can now obtain the metabolic network of a bacterial species (single cell species) of a tissue and even the complete metabolic network of a mammal. Biologists are also moving towards obtaining such networks for all the species present in a microbial community. This will allow us to further investigate the dynamics, the functional profile, and the inter-species reactions that occur. Using the stoichiometry of each reaction, which is always the same in the various species, we convert the metabolic network of an organism to a mathematical model. Thus, the metabolic network becomes an *in silico* model of the knowledge it represents.

In metabolic networks analysis mass and energy are considered to be conserved [182]. As many homeostatic states, that is steady internal conditions [183], are close to steady states (where the production rate of each metabolite equals its consumption rate [184]) we commonly use the latter in metabolic networks analysis.

Stoichiometric coefficients are the number of molecules a biochemical reaction consumes and produces. The coefficients of all the reactions in a network, with m metabolites and n reactions ($m < n$), form the *stoichiometric matrix* $S \in \mathbb{R}^{m \times n}$ [178]. The nullspace of S corresponds to the steady states of the network:

$$S \cdot x = 0, \tag{5.1}$$

where $x \in \mathbb{R}^n$ is the *flux vector* that contains the fluxes of each chemical reaction of the network. Flux is the rate of turnover of molecules through a metabolic pathway.

All physical variables are finite, therefore the flux (and the concentration) is bounded [178]; that is for each coordinate x_i of the x, there are $2n$ constants $x_{ub,i}$ and $x_{lb,i}$ such that $x_{lb,i} \leq x_i \leq x_{ub,i}$, for $i \in [n]$. We derive the constraints from explicit experimental information. In cases where there is no such information, reactions are left unconstrained by setting arbitrary large values to their corresponding bounds according to their reversibility properties; i.e., if a reaction is reversible then its flux might be negative as well [185]. The constraints define a n-dimensional box containing both the steady and the dynamic states of the system. If we intersect that box with the nullspace of S, then we define a polytope that encodes all the possible steady states and their flux distributions [178]. We call it the steady-state *flux space*. Fig. 5.1 illustrates the complete workflow from building a metabolic network to the computation of a flux distribution.

Using the polytopal representation, a commonly used method for the analysis of a metabolic network is Flux Balance Analysis (FBA) [186]. FBA identifies a single optimal flux distribution by optimizing a linear objective function over a polytope [186]. Unfortunately, this is a *biased* method because it depends on the selection of the objective function. To study the global features of a metabolic network we need *unbiased methods*. To obtain an accurate picture of the whole solution space we exploit sampling techniques [187]. If collect a sufficient number of points uniformly distributed in the interior of the polytope, then the biologists can study the properties of certain components of the whole network and deduce significant biological insights [178]. Therefore, efficient sampling tools are of great importance.

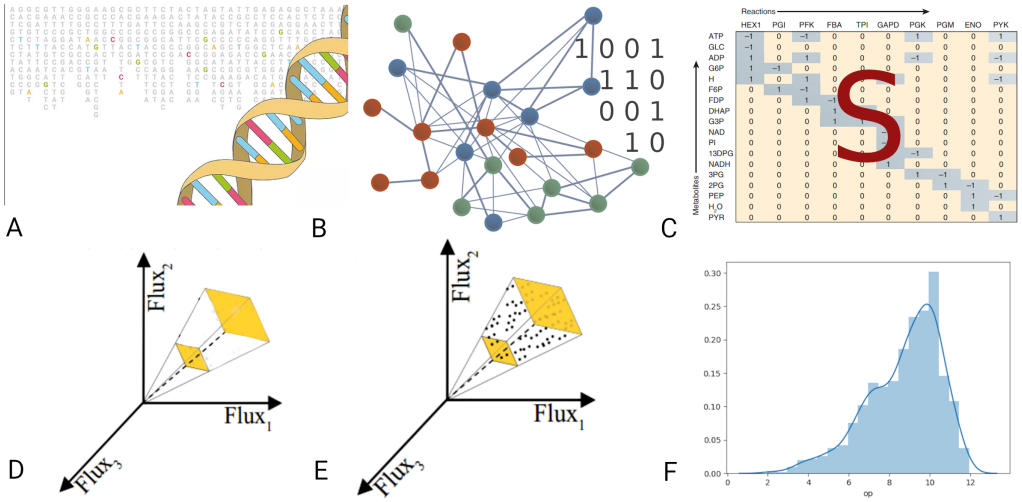

FIGURE 5.1: From DNA sequences to distributions of metabolic fluxes. (A) The genes of an organism provide us with the enzymes that it can potentially produce. Enzymes are like a blueprint for the reactions they can catalyze. (B) Using the enzymes we identify the reactions in the organism. (C) We construct the stoichiometric matrix of the metabolic model. (D) We consider the flux space under different conditions (e.g., steady states); they correspond to polytopes containing flux vectors addressing these conditions. (E) We sample from polytopes that are typically skinny and of high dimension. (F) The distribution of the flux of a reaction provides great insights to biologists.

Metabolic networks through the lens of random sampling

Efficient uniform random sampling on polytopes resulting from metabolic networks is a very challenging task both from the theoretical (algorithmic) and the engineering (implementation) point of view. First, the dimension of the polytopes is of the order of certain thousands. This requires, for example, advanced engineering techniques to cope with memory requirements and to perform linear algebra operations with large matrices; e.g., in Recon3D [2] we compute the null space of a $8\,399 \times 13\,543$ matrix. Second, the polytopes are rather skinny (Sec. 5.1.5); this makes it harder for sampling algorithms to move in the interior of polytopes and calls for novel practical techniques to sample.

There is extended on-going research concerning advanced algorithms and implementations for sampling metabolic networks over the last decades. Markov Chain Monte Carlo algorithms such as Hit-and-Run (HR) [188] have been widely used to address the challenges of sampling. Two variants of HR are the non-Markovian Artificial Centering Hit-and-Run (ACHR) [189] that has been widely used in sampling metabolic models, e.g., [190], and Coordinate Hit-and-Run with Rounding (CHRR) [191]. The latter is part of the cobra toolbox [192], the most commonly used software package for the analysis of metabolic networks. CHRR enables sampling from complex metabolic networks corresponding to the highest dimensional polytopes so far. There are also stochastic formulations where

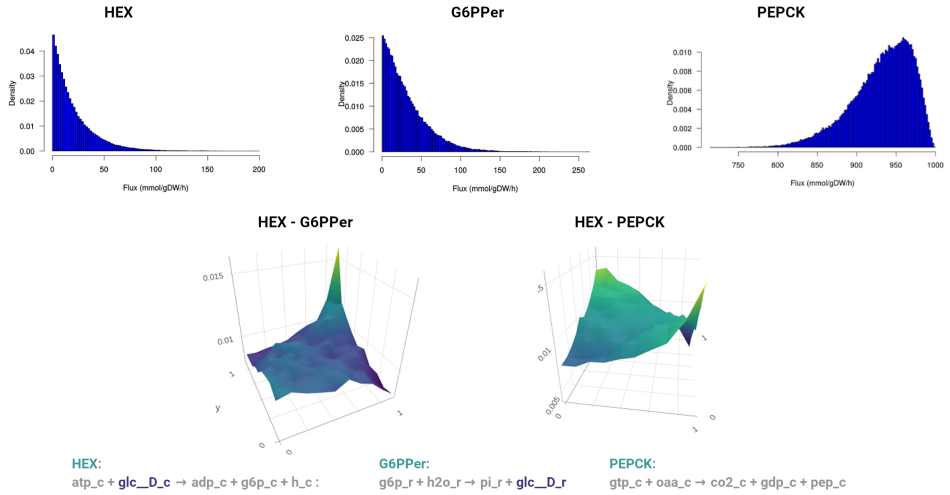

FIGURE 5.2: Flux distributions in the most recent human metabolic network Recon3D [2]. We estimate the flux distributions of the reactions catalyzed by the enzymes Hexokinase (D-Glucose:ATP) (HEX), Glucose-6-Phosphate Phosphatase, Edoplasmic Reticular (G6PPer) and Phosphoenolpyruvate carboxykinase (GTP) (PEPCK). As we sample steady states, the production rate of $glc\_\_D\_c$ should be equal to its consumption rate. Thus, in the corresponding copula, we see a positive dependency between HEX, i.e., the reaction that consumes $glc\_\_D\_c$ and G6PPer, that produces it. Furthermore, the PEPCK reaction operates when there is no $glc\_\_D\_c$ available and does not operate when the latter is present. Thus, in their copula we observe a negative dependency between HEX and PEPCK. A copula is a bivariate probability distribution for which the marginal probability distribution of each variable is uniform. It implies a positive dependency when the mass of the distribution concentrates along the up-diagonal (HEX - G6PPer) and a negative dependency when the mass is concentrated along the down-diagonal (HEX - PEPCK). The bottom line contains the reactions and their stoichiometry.

the inclusion of experimental noise in the model makes it more compatible with the stochastic nature of biological networks [193]. The recent study in [194] offers an overview as well as an experimental comparison of the currently available samplers.

These implementations played a crucial role in actually performing in practice uniform sampling from the flux space. However, they are currently limited to handle polytopes of dimension say $\leq 2\,500$ [194, 191]. This is also the order of magnitude of the most complicated, so far, metabolic network model built, Recon3D [2]. By including 13 543 metabolic reactions and involving 4 140 unique metabolites, Recon3D provides a representation of the 17% of the functionally of annotated human genes. To our knowledge, there is no method that can efficiently handle sampling from the flux space of Recon3D.

Apparently, the dimension of the polytopes will keep rising and not only for the ones corresponding to human metabolic networks. Metabolism governs systems biology at all its levels, including the one of the community. Thus, we are not only interested in sampling a sole metabolic network, even if it has the challenges of the human. Sampling in

polytopes associated to network of networks are the next big thing in metabolic networks analysis and in Systems Biology [195, 196].

Regarding the sampling process, from the theoretical point of view, we are interested in the convergence time, or *mixing time*, of the Markov Chain, or geometric *random walk*, to the target distribution. Given a d-dimensional polytope P, the mixing time of several geometric random walks (e.g., HR or Ball Walk) grows quadratically with respect to the sandwiching ratio R/r of the polytope [197, 198]. Here r and R are the radii of the smallest and largest ball with center the origin that contains, and is contained, in P, respectively; i.e., $rB_d \subseteq P \subseteq RB_d$, where B_d is the unit ball. It is crucial to reduce R/r, i.e., to put P in well a rounded position where $R/r = \widetilde{\mathcal{O}}(\sqrt{d})$; the $\widetilde{\mathcal{O}}(\cdot)$ notation means that we are ignoring polylogarithmic factors. A powerful approach to obtain well roundness is to put P in *near isotropic position*. In general, $K \subset \mathbb{R}^d$ is in isotropic position if the uniform distribution over K is in isotropic position, that is $\mathbb{E}_{X\sim K}[X] = 0$ and $\mathbb{E}_{X\sim K}[X^T X] = I_d$, where I_d is the $d \times d$ identity matrix. Thus, to put a polytope P into isotropic position one has to generate a set of uniform points in its interior and apply to P the transformation that maps the point-set to isotropic position; then iterate this procedure until P is in c-isotropic position [199, 198], for a constant c. In [200] they prove that $\mathcal{O}(d)$ points suffice to achieve 2-isotropic position. Alternatively in [191] they compute the maximum volume ellipsoid in P, they map it to the unit ball, and then apply to P the same transformation. They experimentally show that a few iterations suffice to put P in John's position [201]. Moreover, there are a few algorithmic contributions that combine sampling with distribution isotropization steps, e.g., the multi-point walk [202] and the annealing schedule [203].

An important parameter of a random walk is the walk length, i.e., the number of the intermediate points that a random walk visits before producing a single sample point. The longer the walk length of a random walk is, the smaller the distance of the current distribution to the stationary (target) distribution becomes. For the majority of random walks there are bounds on the walk length to bound the mixing time with respect to a statistical distance. For example, HR generates a sample from a distribution with total variation distance less than ϵ from the target distribution after $\widetilde{\mathcal{O}}(d^3)$ [198] steps, in a well rounded convex body and for log-concave distributions. Similarly, CDHR mixes after a polynomial, in the diameter and the dimension, number of steps [204, 205] for the case of uniform distribution. However, extended practical results have shown that both CDHR and HR converges after $\mathcal{O}(d^2)$ steps [206, 199, 191]. The leading algorithms for uniform polytope sampling are the Riemannian Hamiltonian Monte Carlo sampler [207] and the Vaidya walk [208], with mixing times $\widetilde{\mathcal{O}}(md^{2/3})$ and $\widetilde{\mathcal{O}}(m^{1/2}d^{3/2})$ steps, respectively. However, it is not clear if these random walks can outperform CDHR in practice, because of their high cost per step and numerical instability.

Billiard Walk (BW) [209] is a random walk that employs linear trajectories in a convex body with boundary reflections; alas with an unknown mixing time. The closest guarantees for its mixing time are those of HR and stochastic billiards [210]. Interestingly, [209] shows that, experimentally, BW converges faster than HR for a proper tuning of its parameters. The same conclusion follows from the computation of the volume of zonotopes [211]. It is not known how the sandwiching ratio of P affects the mixing time of BW. Since BW employs reflections on the boundary, we can consider it as a special case of Reflective

Hamiltonian Monte Carlo [212].

For almost all random walks the theoretical bounds on their mixing times are pessimistic and unrealistic for computations. Hence, if we terminate the random walk earlier, we generate samples that are usually highly correlated. There are several *MCMC Convergence Diagnostics* [213] to check if the quality of a sample can provide an accurate approximation of the target distribution. For a dependent sample, a powerful diagnostic is the *Effective Sample Size* (ESS). It is the number of effectively independent draws from the target distribution that the Markov chain is equivalent to. For autocorrelated samples, ESS bounds the uncertainty in estimates [214] and provides information about the quality of the sample. There are several statistical tests to evaluate the quality of a generated sample, e.g., potential scale reduction factor (PSRF) [215], maximum mean discrepancy (MMD) [216], and the uniform tests [217]. Interestingly, the copula representation we employ in Fig. 5.2 to capture the dependence between two fluxes of reactions was also used successfully in a geometric framework to detect financial crises capturing the dependence between portfolio return and volatility [218].

5.1.3 Contribution

We introduce a Multi-phase Monte Carlo Sampling (MMCS) algorithm (Sec. 5.1.4 and Alg. 5.1.4) to sample from a polytope P. In particular, we split the sampling procedure in phases where, starting from P, each phase uses the sample to round the polytope. This improves the efficiency of the random walk in the next phase, see Fig. 5.3. For sampling, we propose an improved variant of Billiard Walk (BW) (Sec. 5.1.4 that enjoys faster arithmetic complexity per step. We also handle efficiently the potential arithmetic inaccuracies near to the boundary, see [212]. We accompany the MMCS algorithm with a powerful MCMC diagnostic, namely the estimation of Effective Sample Size (ESS), to identify a satisfactory convergence to the uniform distribution. However, our method is flexible and we can use any random walk and combination of MCMC diagnostics to decide convergence.

The open-source implementation of our algorithms[1] provides a complete software framework to handle efficiently sampling in metabolic networks. We demonstrate the efficiency of our tools by performing experiments on almost all the metabolic networks that are publicly available and by comparing with the state-of-the-art software packages as *cobra* (Sec. 5.1.7). Our implementation is faster than *cobra* for low dimensional models, with a speed-up that ranges from 10 to 100 times; this gap on running times increases for bigger models (Table 5.1). The quality of the sample our software produces is measured with two widely used diagnostics, i.e., ESS and potential scale reduction factor (PSRF) [215]. The highlight of our method is the ability to sample from the most complicated human metabolic network that is accessible today, namely Recon3D. In Fig. 5.2 we estimate marginal univariate and bivariate flux distributions in Recon3D which validate (a) the quality of the sample by confirming a mutually exclusive pair of biochemical pathways, and that (b) our method indeed generates steady states. In particular, our software can sample $1.44 \cdot 10^5$ points from a 5335-dimensional polytope in a day using modest

[1] https://github.com/GeomScale/volume_approximation/tree/socg21

hardware. This set of points suffices for the majority of systems biology analytics. To our understanding this task is out of reach for existing software. Last, MMCS algorithm is quite general sampling scheme and so it has the potential to address other hard computational problems like multivariate integration and volume estimation of polytopes.

5.1.4 Methods & Implementation

Efficient Billiard walk

The geometric random walk of our choice to sample from a polytope is based on Billiard Walk (BW) [209], which we modify to reduce the per-step cost.

For a polytope $P = \{x \in \mathbb{R}^d \mid Ax \leq b\}$, where $A \in \mathbb{R}^{k\times d}$ and $b \in \mathbb{R}^k$, BW starts from a given point $p_0 \in P$, selects uniformly at random a direction, say v_0, and it moves along the direction of v_0 for length L; it reflects on the boundary if necessary. This results a new point p_1 inside P. We repeat the procedure from p_1. Asymptotically it converges to the uniform distribution over P. The length is $L = -\tau \ln \eta$, where η is a uniform number in $(0,1)$, that is $\eta \sim \mathcal{U}(0,1)$, and τ is a predefined constant. It is useful to set a bound, say ρ, on the number of reflections to avoid computationally hard cases where the trajectory may stuck in corners. In [209] they set $\tau \approx diam(P)$ and $\rho = 10d$. Our choices for τ and ρ depend on a burn-in step that we detail in Sec. 5.1.5.

At each step of BW we compute the intersection point of a ray, say $\ell := \{p + tv,\ t \in \mathbb{R}_+\}$, with the boundary of P, ∂P, and the normal vector of the tangent plane at the intersection point. The inner vector of the facet that the intersection point belongs to is a row of A. To compute the point $\partial P \cap \ell$ where the first reflection of a BW step takes place, we solve the following m linear equations

$$a_j^T(p_0 + t_j v_0) = b_j \Rightarrow t_j = (b_j - a_j^T p_0)/a_j^T v_0,\ j \in [k], \tag{5.2}$$

and keep the smallest positive t_j; a_j is the j-th row of the matrix A. We solve each equation in $\mathcal{O}(d)$ operations and so the overall complexity is $\mathcal{O}(dk)$. A straightforward approach for BW would consider that each reflection costs $\mathcal{O}(kd)$ and thus the per step cost is $\mathcal{O}(\rho kd)$. However, our improved version performs more efficiently both *point* and *direction updates* by storing computations from the previous iteration combined with a preprocessing step. The preprocessing step involves the normal vectors of the facets, that takes $m^2 d$ operations, and the amortized per-step complexity of BW becomes $\mathcal{O}((\rho + d)k)$.

Lemma 1. *The amortized per step complexity of BW is $\mathcal{O}((\rho + d)k)$ after a preprocessing step that takes $\mathcal{O}(k^2 d)$ operations, where ρ is the maximum number of reflections per step.*

Proof. The first reflection of a Billiard Walk step costs $O(kd)$. During its computation, we store all the values of the inner products $a_j^T x_0$ and $a_j^T u_0$. At the reflection $i > 0$, we start

from a point x_i and the solutions of the corresponding linear equations are

$$a_j^T(p_i + t_j u_i) = b_j \Rightarrow$$

$$a_j^T(p_{i-1} + t_{i-1}u_{i-1}) + t_j a_j^T(u_{i-1} - 2(u_{i-1}^T a_r)a_r) = b_j \Rightarrow$$

$$t_j = \frac{b_j - a_j^T(p_{i-1} + t_{i-1}u_{i-1})}{a_j^T(u_{i-1} - 2(u_{i-1}^T a_r)a_r)}, \text{ for } j \in [k],$$

$$\text{and } u_{i+1} = u_i - 2(u_i^T a_l)a_l, \tag{5.3}$$

where a_r, a_l are the normal vectors of the facets that ℓ hits at reflection $i-1$ and i respectively, and t_{i-1} the solution of the reflection $i-1$. The index l of the normal a_l corresponds to the equation with the smallest positive t_j in (5.1.4). We solve each of the equations in (5.3) in $O(1)$ based on our bookkeeping from the previous reflection. We also store the inner product $u_i^T a_l$ in (5.3) from the previous reflection. After computing all $a_i^T a_j$ as a preprocessing step, which takes $k^2 d$ operations, the total per-step cost of Billiard Walk is $O((d+\rho)k)$. □

The use of floating point arithmetic could result to points outside P due to rounding errors when computing boundary points. To avoid this, when we compute the roots in Equation (5.2) we exclude the facet that the ray hit in the previous reflection.

At each step of Billiard Walk, we compute the intersection point of a ray, say $\ell := \{p + tu,\ t \in \mathbb{R}_+\}$, with the boundary of P, ∂P, and the normal vector of the tangent plane of P at the intersection point. The inner vector of the facet that the intersection point belongs to is a row of A. To compute the point $\partial P \cap \ell$ where the first reflection of a Billiard Walk step takes place we need to compute the intersection of ℓ with all the hyperplanes that define the facets of P. This corresponds to solve (independently) the following m linear equations

$$a_j^T(p_0 + t_j u_0) = b_j \Rightarrow t_j = (b_j - a_j^T p_0)/a_j^T u_0,\ j \in [k], \tag{5.4}$$

and keep the smallest positive t_j; a_j is the j-th row of the matrix A. We solve each equation in $\mathcal{O}(d)$ operations and so the overall complexity is $\mathcal{O}(dk)$, where k is the number of rows of A and thus an upper bound on the number of facets of P. A straightforward approach for Billiard Walk would consider that each reflection costs $\mathcal{O}(kd)$ and thus the per step cost is $\mathcal{O}(\rho kd)$. However, our improved version performs more efficiently both *point* and *direction updates* in pseudo-code by storing some computations from the previous iteration combined with a preprocessing step. The preprocessing step involves the normal vectors of the facets and takes $k^2 d$ operations. So the amortized per-step complexity of Billiard Walk becomes $\mathcal{O}((\rho + d)k)$. The pseudo-code appear in Algorithm 5.1.4.

Algorithm 1: Billiard Walk(P, p, ρ, τ, W)

Require: polytope P; point $p \in P$; upper bound on the number of reflections ρ; parameter τ to adjust the length of the trajectory; walk length W.
Ensure: a point in P (uniformly distributed in P).
for $j = 1, \ldots, W$ **do**
$L \leftarrow -\tau \ln \eta$; $\eta \sim \mathcal{U}(0,1)$ {*length of the trajectory*} $i \leftarrow 0$ {*current number of reflections*} $p_0 \leftarrow p$ {*initial point of the step*} pick a uniform vector u_0 from the unit sphere {*initial direction*}
while $i \leq \rho$ **do**
$\ell \leftarrow \{p_i + tu_i, 0 \leq t \leq L\}$ {*this is a segment*}
if $\partial P \cap \ell = O$ **then**
$p_{i+1} \leftarrow p_i + Lu_i$ **break**
end if
$p_{i+1} \leftarrow \partial P \cap \ell$; {*point update*}
the inner vector, s, of the tangent plane at p,
s.t. $||s|| = 1$, $L \leftarrow L - |P \cap \ell|$, $u_{i+1} \leftarrow u_i - 2(u_i^T s)s$ {*direction update*}
$i \leftarrow i + 1$
end while
if $i = \rho$ **then**
$p \leftarrow p_0$
else
$p \leftarrow p_i$
end if
end for
return p

Multiphase Monte Carlo Sampling algorithm

To sample steady states in the flux space of a metabolic network, with m metabolites and n reactions, we introduce a Multiphase Monte Carlo Sampling (MMCS) algorithm; it is multiphase because it consists of a sequence of sampling phases.

Let $S \in \mathbb{R}^{m \times n}$ be the stoichiometric matrix and x_{lb}, $x_{ub} \in \mathbb{R}^n$ bounds on the fluxes. The flux space is the bounded convex polytope

$$\text{FS} := \{x \in \mathbb{R}^n \mid Sx = 0,\ x_{lb} \leq x \leq x_{ub}\} \subset \mathbb{R}^n. \tag{5.5}$$

The dimension, d, of FS is smaller than the dimension of the ambient space; that is $d \leq n$. To work with a full dimensional polytope we restrict the box induced by the inequalities $x_{lb} \leq x \leq x_{ub}$ to the null space of S. Let the H-representation of the box be $\left\{x \in \mathbb{R}^n \,\middle|\, \begin{pmatrix} I_n \\ -I_n \end{pmatrix} x \leq \begin{pmatrix} x_{ub} \\ x_{lb} \end{pmatrix}\right\}$, where I_n is the $n \times n$ identity matrix, and let $N \in \mathbb{R}^{n \times d}$ be the matrix of the null space of S, that is $SN = 0_{m \times d}$. Then $P = \{x \in \mathbb{R}^d \mid Ax \leq b\}$, where $A = \begin{pmatrix} I_n N \\ -I_n N \end{pmatrix}$ and $b = \begin{pmatrix} x_{ub} \\ x_{lb} \end{pmatrix} N$, is a full dimensional polytope (in $\mathbb{R}^d$). After we sample

(uniformly) points from P, we transform them to uniformly distributed points (that is steady states) in FS by applying the linear map induced by N.

MMCS generates, in a sequence of sampling phases, a set of points, that is almost equivalent to n independent uniformly distributed points in P, where n is given. At each phase, it employs Billiard Walk (Section 5.1.4) to sample approximate uniformly distributed points, rounding to speedup sampling, and uses the Effective Sample Size (ESS) diagnostic to decide termination. The pseudo-code of the algorithm appears in Alg. 5.1.4.

Overview. Initially we set $P_0 = P$.

At each phase $i \geq 0$ we sample at most λ points from P_i. We generate them in chunks; we also call them *chain* of sampling points. Each chain contains at most l points (for simplicity consider $l = \mathcal{O}(1)$). To generate the points in each chain we employ BW, starting from a point inside P_i; the starting point is different for each chain. We repeat this procedure until the total number of samples in P_i reaches the maximum number λ; we need $\frac{\lambda}{l}$ chains. To compute a starting point for a chain, we pick a point uniformly at random in the Chebychev ball of P_i and we perform $\mathcal{O}(\sqrt{d})$ burn-in BW steps to obtain a warm start.

After we have generated λ sample points we perform a rounding step on P_i to obtain the polytope of the next phase, P_{i+1}. We compute a linear transformation, T_i, that puts the sample into isotropic position and then $P_{i+1} = T_i(P_i)$. The efficiency of BW improves from one phase to the next one because the sandwiching ratio decreases and so the average number of reflections decreases and thus the convergence to the uniform distribution accelerates (Section 5.1.7). That is we obtain faster a sample of better quality. Finally, the (product of the) inverse transformations maps the samples to $P_0 = P$. Fig. 5.3 depicts the procedure.

Termination. There are no bounds on the mixing time of BW [209], hence for termination we rely on ESS. MMCS terminates when the minimum ESS among all the univariate marginals is larger than a requested value. We chose the marginal distributions (of each flux) because they are essential for systems biologists, see [219] for a typical example. In particular, after we generate a chain, the algorithm updates the ESS of each univariate marginal to take into account all the points that we have sampled in P_i, including the newly generated chain. We keep the minimum, say n_i, among all marginal ESS values. If $\sum_{j=0}^{i} n_j$ becomes larger than n before the total number of samples in P_i reaches the upper bound λ, then MMCS terminates. Otherwise, we proceed to the next phase. In summary, MMCS terminates when the sum of the minimum marginal ESS values of each phase reaches n.

Rounding step. This step is motivated by the theoretical result in [200] and the rounding algorithms [198, 199]. We apply the linear transformation T_i to P_i so that the sandwiching ratio of P_{i+1} is smaller than that of P_i. To find the suitable T_i we compute the SVD decomposition of the matrix that contains the sample row-wise [220].

Updating the Effective Sample Size. The effective sample size of a sample of points generated by a process with autocorrelations ρ_t at lag t is function (actually an infinite series) in the ρ_t's; its exact value is unknown. Following [214], we efficiently compute ESS employing a finite sum of monotone estimators $\hat{\rho}_t$ of the autocorrelation at lag

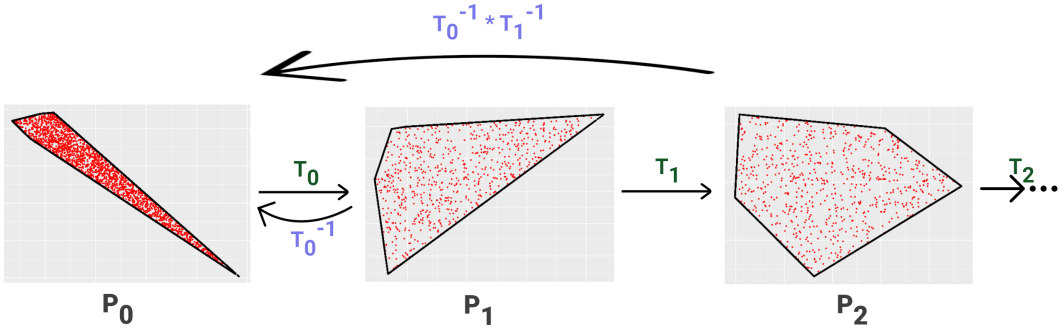

FIGURE 5.3: An illustration of our Multiphase Monte Carlo Sampling algorithm. The method is given an integer n and starts at phase $i = 0$ sampling from P_0. In each phase it samples a maximum number of points λ. If the sum of Effective Sample Size in each phase becomes larger than n before the total number of samples in P_i reaches λ then the algorithm terminates. Otherwise, we proceed to a new phase. We map back to P_0 all the generated samples of each phase.

t, by exploiting Fast Fourier Transform. Furthermore, given M chains of samples, the autocorrelation estimator $\hat{\rho}_t$ is given by, $\hat{\rho}_t = 1 - \frac{C - \frac{1}{M}\sum_{i=1}^{M} \hat{\rho}_{t,i}}{B}$, where C and B are the within-sample variance estimate and the multi-chain variance estimate given in [215] and $\hat{\rho}_{t,i}$ is an estimator of the autocorrelation of the i-th chain at lag t. To update the ESS, for every new chain of points the algorithm generates, we compute $\hat{\rho}_{t,i}$. Then, using Welford's algorithm we update the average of the estimators of autocorrelation at lag t, as well as the between-chain variance and the within-sample variance estimators given in [215]. Finally, we update the ESS using these estimators.

To update the ESS, for every new chain of points the algorithm generates, we compute the estimator of its autocorrelation. Then, using Welford's algorithm we update the average of the estimators of autocorrelation at lag t, as well as the between-chain variance and the within-sample variance estimators [215]. Finally, we update the ESS using these estimators.

Lemma 2 (Complexity of MMCS per phase)**.** *Let $P = \{x \in \mathbb{R}^d \mid Ax \leq b\}$, where $A \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$, be a full dimensional polytope in $\mathbb{R}^d$. To sample n points (approximately) uniformly distributed in P, MMCS (Algorithm 5.1.4) performs $\mathcal{O}(W(\rho + d)k\lambda + \lambda^2 d + d^3)$ arithmetic operations per phase, where W is the walk length of Billiard Walk, ρ is an upper bound on the number of reflections, and λ and upper bound on the points generated at each phase.*

Proof. The cost per step of Billiard Walk is $O((\rho + d)k)$. In each phase we generate, using Billiard Walk, at most λ points with walk length W. Thus, the cost to generate these points is $O(W(\rho + d)k\lambda)$.

To compute the starting point of each chain the algorithm picks a random point uniformly distributed in the Chebychev ball of P and performs $\mathcal{O}(1)$ Billiard Walk steps starting from it. The former takes $\mathcal{O}(d)$ operations and the latter takes $\mathcal{O}(W(\rho + d)k)$ operations. The total number of chains is $\mathcal{O}(\lambda/l) = \mathcal{O}(\lambda)$, as $l = \mathcal{O}(1)$. Thus, the total cost

Algorithm 2: Multiphase Monte Carlo Sampling$(P, n, l, \lambda, \rho, \tau, W)$

Require: A full dimensional polytope $P \in \mathbb{R}^d$;
requested effectiveness $n \in \mathbb{N}$ (number of sampled points);
l length of each chain;
λ upper bound of the number of generated points in each phase λ;
upper bound on the number of reflections ρ;
parameter τ to adjust the length of the trajectory; walk length W.

Ensure: a set n of approximate uniformly distributed points $S \in P$

Set $P_0 \leftarrow P$, $sum\_ess \leftarrow 0$, $S \leftarrow \emptyset$, $i \leftarrow 0$, $T_0 = I_d$
while $sum\_ess < n$ **do**
  $sum\_point\_phase \leftarrow 0$, $U \leftarrow \emptyset$
  while $sum\_point\_phase < \lambda$; **do**
    Set $Q \leftarrow \emptyset$; Generate a starting point $q_0 \in P_i$;
    for $j = 1, \ldots, l$ **do**
      $q_j \leftarrow$Billiard_Walk$(P_i, q_{j-1}, \rho, \tau, W)$, Store the point q_j to the set Q
    end for
    $S \leftarrow S \cup T_i^{-1}(Q)$, $U \leftarrow U \cup Q$, $sum\_point\_phase \leftarrow sum\_point\_phase + l$ Update ESS n_i of this phase
    if $sum\_ess + n_i \geq n$ **then**
      break
    end if
  end while
  $sum\_ess \leftarrow sum\_ess + n_i$, Compute T such that $T(U)$ is in isotropic position, $P_{i+1} \leftarrow T(P_i)$, $T_{i+1} \leftarrow T_i \circ T$, $i \leftarrow i + 1$
end while
return S

to generate all the starting points is $\mathcal{O}(d\lambda + W(\rho + d)k\lambda)$. The update of ESS for each univariate marginal requires $\mathcal{O}(1)$ operations, since $l = \mathcal{O}(1)$.

If the termination criterion has not been met after generating λ points, then the algorithm computes a linear transformation to put the set of points to isotropic position. We do this by computing the SVD decomposition of the matrix that contains the set of points row-wise. This corresponds to an SVD of a $\lambda \times d$ matrix and takes $O(\lambda^2 d + d^3)$ operations [221]. □

In Section 5.1.5 we discuss how to tune the parameters of MMCS to make it more efficient in practice. We also comment on the (practical) complexity of each phase, based on the tuning.

5.1.5 Results

Implementation and Experiments

This section presents the implementation of our approach and the tuning of various parameters. We present experiments in an extended set of BiGG models [222], including the most complex metabolic networks, the human Recon2D [1] and Recon3D [2]. We end up to sample from polytopes of thousands of dimensions and show that our method can estimate precisely the flux distributions. We analyze various aspects of our method such as the run-time, the efficiency, and the quality of the output.

We compare against the state-of-the-art software for the analysis of metabolic networks, which is the Matlab toolbox of cobra [192]. Our implementation for low dimensional networks is two orders of magnitude faster than cobra. As the dimension grows, this gap on the run-time increases. The workflow of cobra for sampling first performs a rounding step and then samples using Coordinate Directions Hit-and-Run (CDHR).

In [223] they provide a C++ implementation of the sampling method that cobra uses and they show that their implementation is approximately 6 times faster than cobra. Nevertheless, we choose to compare against cobra, since it additionally provides efficient preprocessing methods that are crucial for the experiments, and give an implicit comparison with [223].

The fast mixing of billiard walk allow us to use all the generated samples to approximate each flux distribution and so we compute a better flux distribution estimation. To estimate each marginal flux distribution, using the samples, we exploit Gaussian kernel density estimation. This is a non-parametric way to estimate the probability density function of a random variable. For more details we refer to [224].

We provide a complete open-source software framework to handle big metabolic networks. The framework loads a metabolic model in some standard file formats (e.g., mat and json files) and performs an analysis of the model, e.g., it estimates the marginal distributions of a given reaction flux. All the results are reproducible using our publicly available code

The core of our implementation is in C++ to optimize performance while the user interface is implemented in R. The package employs [225] for linear algebra, number generation, [226], an open-source package for high dimensional sampling and volume approximation.

All experiments were performed on a PC with Intel Core i7-6700 3.40GHz × 8 CPU and 32GB RAM. In the sequel, MMCS refers to our implementation.

5.1.6 Parameter tuning for practical performance

We give details on how we tune the various parameters presented in Section 5.1.4 in our implementation.

Parameters of Billiard Walk: To employ Billiard Walk (Section 5.1.4), we have to efficiently select values for the parameter τ that controls the length of the trajectory in each step, for the maximum number of reflections per step ρ, and for the walk length W of the random walk. We have experimentally found that if we set $W = 1$, then the empirical distribution converges faster to the uniform distribution. Thus, we get a higher ESS faster

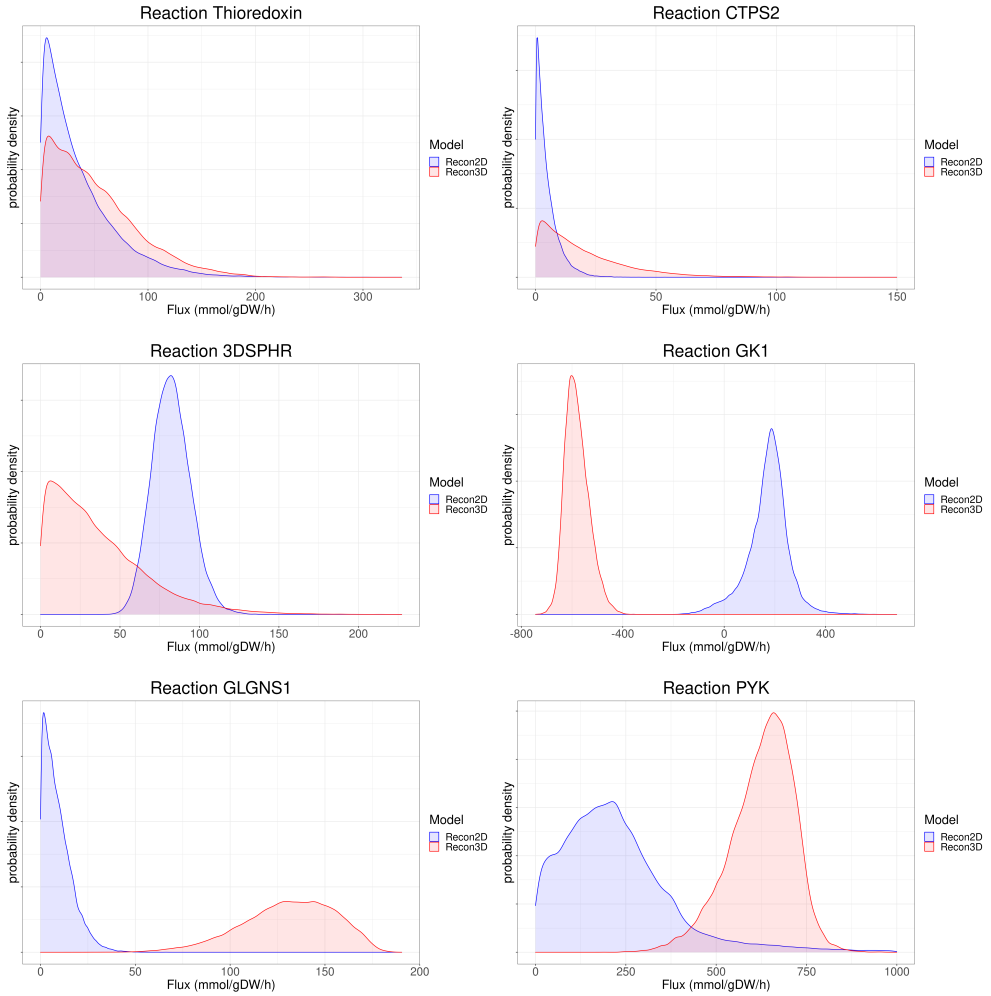

FIGURE 5.4: Estimation, using our tools, of the marginal distribution of 6 reaction fluxes in two constraint-based model of Homo Sapiens metabolism, namely Recon2D [1] (blue color) and Recon3D [2] (red color). In the case of GK1 we observe how the flux distribution of a reaction may change once the direction of the reaction changes.

than the case of $W > 1$. To set τ in phase i, first we set $\tau = 6\sqrt{d}r$, where r is the radius of the Chebychev ball of P_i. Next, we start from the center of the Chebychev ball, we perform $100 + 4\sqrt{d}$ Billiard Walk steps, and we store all the points in a set Q. Then we set $\tau = \max\{\max_{q \in Q}\{||q - p||_2\}, 6\sqrt{d}r\}$. For the maximum number of reflections we have found experimentally that $\rho = 100d$ is violated in less than 0.1% of the total number of Billiard Walk steps in our experiments.

Rounding step: In each phase i of our method, if the minimum value of ESS among all the marginals has not reached the requested threshold, then we use the generated sample to perform a rounding step by mapping the points to isotropic position. After we compute the SVD decomposition of (the matrix corresponding to) the point-set, we

				MMCS		cobra	
model	m	n	d	Time (sec)	N	Time (sec)	N
e_coli_core	72	95	24	6.50e-01	3.40e+03 (8)	7.20e+01	4.61e+06
iLJ478	570	652	59	9.00e+00	5.40e+03 (5)	4.54e+02	2.79e+07
iSB619	655	743	83	1.70e+01	8.20e+03 (5)	9.56e+02	5.51e+07
iHN637	698	785	88	2.00e+01	6.80e+03 (4)	1.03e+03	6.19e+07
iJN678	795	863	91	2.50e+01	8.10e+03 (4)	1.17e+03	6.62e+07
iNF517	650	754	92	1.70e+01	6.20e+03 (4)	1.33e+03	6.77e+07
iJN746	907	1054	116	5.70e+01	8.70e+03 (3)	2.22e+03	1.07e+08
iAB_RBC_283	342	469	130	5.20e+01	1.07e+04 (5)	7.85e+03	4.05e+08
iJR904	761	1075	227	2.98e+02	1.62e+04 (4)	8.81e+03	4.12e+08
iAT_PLT_636	738	1008	289	3.25e+02	1.04e+04 (2)	1.73e+04	6.68e+08
iSDY_1059	1888	2539	509	2.813e+03	2.31e+04 (3)	6.66e+04	2.07e+09
iAF1260	1668	2382	516	6.84e+03	5.33e+04 (6)	7.04e+04	2.13e+09
iEC1344_C	1934	2726	578	4.86e+03	3.95e+04 (4)	9.42e+04	2.67e+09
iJO1366	1805	2583	582	6.02e+03	5.14e+04 (5)	9.99e+04	2.71e+09
iBWG_1329	1949	2741	609	3.06e+03	4.22e+04 (4)	1.05e+05	2.97e+09
iML1515	1877	2712	633	4.65e+03	5.65e+04 (5)	1.15e+05	3.21e+09
Recon1	2766	3741	931	8.09e+03	1.94e+04 (2)	3.20e+05	6.93e+09
Recon2D	5063	7440	2430	2.48e+04	5.44e+04 (2)	~ 140 days	$1.57e$+11
Recon3D	8399	13543	5335	1.03e+05	1.44e+05 (2)	–	–

TABLE 5.1: Several, 17, metabolic networks from [222]; also Recon2D and Recon3D from [227]. The semantics of the tables are as follows: (m) the number of Metabolites, (n) the number of Reactions, (d) the dimension of the polytope; (N) is the total number of sampled points × walk length; for MMCS we stop when the sum of the minimum value of ESS among all the univariate marginals in each phase is 1 000 (we report the number of phases in parenthesis); for cobra we set the walk length to $8d^2$ and $1.57e$+08 for Recon2D following [191], sample at least 1 000 points and stop when all marginals have PSRF < 1.1; the run-time of cobra for Recon2D is an estimation of the sequential time and we report it to have a rough comparison with our implementation.

rescale the singular values so that the smallest one is 1; this is to improve numerical stability as suggested in [199].

We have also experimentally found that to improve the roundness from phase to phase, it suffices to set the maximum number of Billiard Walk points per phase to $\lambda = 20d$, where d is the dimension of the polytope. When, in any phase, the ratio between the maximum and the minimum singular value is smaller than 3, then we do not perform any new rounding step. In this case, we stay on the current phase until we reach the requested ESS value.

Remark 3. *Given the stoichiometric matrix $S \in \mathbb{R}^{m\times n}$ of a metabolic network with flux bounds $v_{lb} \le v \le v_{ub}$, the total number of operations per phase that our implementation MMCS (Algorithm 5.1.4) performs, according the parameterization given in this section, is $\mathcal{O}(nd^2)$, where d is the dimension of the nullspace of S and n is the number of reactions occur in the metabolic network.*

5.1.7 Experiments

We test and evaluate our software on 17 models from the BIGG database [222] as well as Recon2D and Recon3D from [227]. In particular, we sample from models that correspond to polytopes of dimension less than 100; the simplest model in this setting is the well known bacteria *Escherichia Coli*. We also sample from models that correspond to polytopes of dimension a few thousands; this is the case for Recon2D and Recon3D. We do not employ parallelism for any implementation, thus we report only sequential running times.

We assess the quality of our results by employing both the Effective Sample Size (ESS) and the potential scale reduction factor (PSRF) [215]. In particular, we compute the PSRF for each univariate marginal of the sample that MMCS outputs. Following [215], a convergence is satisfying according to PSRF when all the marginals have PSRF smaller than 1.1.

In Table 5.1, we report the results of MMCS and cobra. For cobra, we report only the run-time of the sampling phase (we do not add to it the preprocessing time). We run MMCS until we get a value of ESS equal to 1 000; i.e. we stop when the sum over all phases of the minimum values of ESS among all the marginals is larger than 1 000. All the marginals of the MMCS samples reported in Table 5.1 have PSRF < 1.1. This is a strong statistical evidence on the quality of the generated sample.

The histograms in Figure 5.4 illustrate an approximation for the flux distribution of 6 reaction fluxes in Recon2D and Recon3D, respectively. Notice the difference in estimated densities due to the stoichiometric matrix update from Recon2D to Recon3D. The marginal flux distribution of reaction Thioredoxin in Recon2D was estimated also in [191] and used as an evidence for the quality of the sample. In Figure 5.2, we employ the copula representation to capture the dependency between two fluxes of reactions and confirm a mutually exclusive pair of biochemical pathways.

Comparing runtime performance, MMCS is one or two orders of magnitude faster than cobra and this gap becomes much larger for higher dimensional models such as Recon2D and Recon3D. Considering the experiments reported in [223], they report the run-time of CDHR for each model until it generates a sample with PSFR 1.2; for Recon3D they report ~ 1 day. Interestingly, for Recon3D, MMCS achieves PSRF 1.2 after ~ 1 hour while reach PSRF 1.1 after ~ 1 day.

For some models –we report them in Table 5.2– we introduce a further improvement to obtain a better convergence. If there is a marginal in the generated sample from MMCS that has a PSRF larger than 1.1, then we do not take into account the k first phases, starting with $k = 1$ until we get both ESS equal to 1 000 and all the PSRF values smaller than 1.1 for all the marginals. By "we do not take into account" we mean that we neither store the generated sample –for the first k phases– nor we sum up its ESS to the overall ESS considered for termination by MMCS. Note that for these models it is not practical to repeat MMCS runs for different k until we get the required PSRF value. We can obtain the final results –reported in Tables 5.1– in one pass. We simply drop a phase when the ESS reaches the requested value but the PSRF is not smaller than 1.1 for all the marginals. In Table 5.2, we separately report the MMCS runs for different k just for performance analysis reasons.

model	k	Time (sec)	PSRF < 1.1	M	N
iAF1260	0	6955	41%	6	56100
	1	6943	56%	6	54100
	2	6890	76%	6	55200
	3	6867	95%	6	53200
	4	6840	100%	6	53300
iBWG_1329	0	3067	50%	4	42100
	1	3189	97%	5	48800
	2	4652	100%	5	56500
iEC1344	0	4845	77%	4	41100
	1	4721	96%	4	42500
	2	4682	100%	4	39500
iJO1366	0	3708	66%	5	51500
	1	6022	100%	5	51400

TABLE 5.2: During our experiments we do not take into account the sample of the k first phases, thus we do not also count the value of the Effective Sample Size (ESS) in these phases, before we start storing the generated sample and sum up the ESS of each phase. In all cases MMCS stops when the sum of ESS reaches 1000. For each case we report the total run-time, the percentage of the marginals that have PSRF smaller than 1.1, the total number of phases (M) needed (including the k first phases), and the total number of Billiard Walk steps (N), including those performed in the k first phases.

Interestingly, the total number of Billiard Walk steps –and consequently the run-time– does not increase as k increases in Table 5.2. This means that the performance of our method improves for these models when we do not take into account the k first phases of MMCS. This happens because the performance of Billiard Walk improves as the polytope becomes more rounded from phase to phase.

In Table 5.3, we analyze the performance of Billiard Walk for the model iAF1260. We sample $20d$ points per phase with walk length equal to 1 and we report the average number of reflections, the ESS, the run-time, and the ratio $\sigma_{\max}/\sigma_{\min}$ per phase. The latter is the ratio of the maximum over the minimum singular value of the point-set. The larger this ratio is the more skinny the polytope of the corresponding phase is. As the method progresses from the first to the last phase, the average number of reflections and the run-time decrease and the ESS increases. This means that as the polytope becomes more rounded from phase to phase, the Billiard Walk step becomes faster and the generated sample has better quality. This explains why the total run-time does not increase when we do not take into account the first k phases: the initial phases are slow and they contribute poorly to the quality of the final sample; the last phases are fast and contribute with more accurate samples.

5.2 Conclusions and future work

We propose a novel method for sampling that can sample from a convex polytope in a few thousands of dimensions within a day on modest hardware. This way, we are able, for the first time, to perform accurate sampling from the latest human metabolic network, Recon3D.

Sampling from iAF1260				
Phase	Avg. #reflections	ESS	$\frac{\sigma_{\max}}{\sigma_{\min}}$	Time (sec)
1st	7819	67	43459	2271
2nd	4909	68	922	1631
3rd	3863	77	582	1278
4th	3198	71	360	1080
5th	1300	592	29	454
6th	1187	4821	3.5	417
7th	1181	4567	2.8	415

TABLE 5.3: We sample $20d = 10320$ points per phase with Billiard Walk and walk length equal to 1, where $d = 516$ is the dimension of the corresponding polytope. For each phase we report the average number of reflections per BIlliard Walk step, the minimum value of Effective Sample Size among all the univariate marginals, the ratio between the maximum and the minimum singular value of the SVD decomposition of the generated sample, and the run-time.

Regarding future work, parallelism could lead to a speedup in the run-time of our method as the algorithm is rather straightforward to parallelize. An additional improvement would be to exploit the sparsity of the stoichiometric matrix S and sample directly from the low dimensional polytope in $\mathbb{R}^n$ without projecting to a lower dimensional space.

Moreover, our method could be extended to any log-concave distribution restricted to the flux space and combined with bayesian metabolic flux analysis, to sample from multivariate, possibly multi-modal target distribution [228] addressing multiple challenges of the method from the biological point of view (e.g., unrealistic assumptions, uncertainty etc.). Last but not least, flux sampling in metabolic models built out from multiple metabolic networks, e.g., representing a microbial community, could also lead to important biological insights.

Chapter 6

An overview of the computational requirements & solutions in microbial ecology

6.1 0s and 1s in marine molecular research: a regional HPC perspective[1]

Citation: Zafeiropoulos, H., Gioti, A., Ninidakis, S., Potirakis, A., Paragkamian, S., ... & Pafilis, E. (2021). 0s and 1s in marine molecular research: a regional HPC perspective. GigaScience, 10(8), giab053, doi: 10.1093/gigascience/giab053

6.1.1 Abstract

High-performance computing (HPC) systems have become indispensable for modern marine research, providing support to an increasing number and diversity of users. Pairing with the impetus offered by high-throughput methods to key areas such as non-model organism studies, their operation continuously evolves to meet the corresponding computational challenges.

Here, we present a Tier 2 (regional) HPC facility, operating for over a decade at the Institute of Marine Biology, Biotechnology, and Aquaculture of the Hellenic Centre for Marine Research in Greece. Strategic choices made in design and upgrades aimed to strike a balance between depth (the need for a few high-memory nodes) and breadth (a number of slimmer nodes), as dictated by the idiosyncrasy of the supported research. Qualitative computational requirement analysis of the latter revealed the diversity of marine fields, methods, and approaches adopted to translate data into knowledge. In addition, hardware and software architectures, usage statistics, policy, and user management aspects of the facility are presented. Drawing upon the last decade's experience from the different levels of operation of the Institute of Marine Biology, Biotechnology, and Aquaculture HPC facility, a number of lessons are presented; these have contributed to the facility's future

[1] For author contributions, please refer to the relevant section. Modified version of the published review.

directions in light of emerging distribution technologies (e.g., containers) and Research Infrastructure evolution. In combination with detailed knowledge of the facility usage and its upcoming upgrade, future collaborations in marine research and beyond are envisioned.

6.1.2 Introduction

The ubiquitous marine environments (more than 70% of the global surface [229]) mold Earth's conditions to a great extent. The interconnected abiotic [4] and biotic factors (from bacteria [4] to megafauna [230]), shape biogeochemical cycles [231] and climates [232, 233] from local to global scales. In addition, marine systems have high socio-economic value [234] as an essential source of food and by supporting renewable energy and transport, among other services [235]. The study of marine environments involves a series of disciplines (scientific fields): from Biodiversity [236] and Oceanography to (eco)systems biology [237] and from Biotechnology [238] to Aquaculture [239].

To shed light on the evolutionary history of (commercially important) marine species [240], as well as on how invasive species respond and adapt to novel environments [241], the analysis of their genetic stock structure is fundamental [242]. Similarly, biodiversity assessment is essential to elucidate ecosystem functioning [243] and to identify taxa with potential for bioprospecting applications [244]. Furthermore, systems biology approaches provide both theoretical and technical backgrounds in which integrative analyses flourish [245]. However, conventional methods do not offer the information needed to explore the aforementioned scientific topics.

High-throughput sequencing (HTS) and sister methods have launched a new era in many biological disciplines [246, 247]. These technologies allowed access to the genetic, transcript, protein, and metabolite repertoire [248] of studied taxa or populations, and facilitated the analysis of organism-environment interactions in communities and ecosystems [249]. Whole-genome sequencing and whole-transcriptome sequencing approaches provide valuable information for the study of non-model taxa [250]. This information can be further enriched by genotyping-by-sequencing approaches, such as restriction site-associated DNA sequencing [251], or by investigating gene expression dynamics through Differential Expression (DE) analyses [252]. Moving from single species to assemblages, molecular-based identification and functional profiling of communities has become available through marker (metabarcoding), genome (metagenomics), or transcriptome (metatranscriptomics) sequencing from environmental samples [253]. To a great extent, these methods address the problem of how to produce and get access to the information on different biological systems and molecules.

These 0s and 1s of information (i.e., the data) come along with challenges regarding their management, analysis, and integration [254]. The computational requirements for these tasks exceed the capacity of a standard laptop/desktop by far, owing to the sheer volume of the data and to the computational complexity of the bioinformatic algorithms employed for their analysis. For example, building the de novo genome assembly of a non-model Eukaryote may require algorithms of nondeterministic polynomial time complexity. This analysis can reach up to several hundreds or thousands of GB of memory

(RAM) [255]. Hence, the challenges of how to exploit all these data and how to transform data into knowledge set the present framework in biological research [256, 257].

To address these computational challenges, the use of high-performance computing (HPC) systems has become essential in life sciences and systems biology [258]. HPC is the scientific field that aims at the optimal incorporation of technology, methodology, and the application thereof to achieve “the greatest computing capability possible at any point in time and technology" [259]. Such systems range from a small number to several thousands of interconnected computers (compute nodes). According to the Partnership for Advanced Computing in Europe, the European HPC facilities are categorized as: (i) European Centres (Tier 0), (ii) national centers (Tier 1), and (iii) regional centers (Tier 2) [33] [260]. As the Partnership for Advanced Computing in Europe highlights, "computing drives science and science drives computing" in a great range of scientific fields, from the endeavor to maintain a sustainable Earth to efforts to expand the frontiers in our understanding of the universe [261]. On top of the heavy computational requirements, biological analyses come with a series of other practical issues that often affect the bioinformatics-oriented HPC systems.

Researchers with purely biological backgrounds often lack the coding skills or even the familiarity required for working with Command Line Interfaces [261]. Virtual Research Environments are web-based e-service platforms that are particularly useful for researchers lacking expertise and/or computing resources [262]. Another common issue is that most analyses include a great number of steps, with the software used in each of these having equally numerous dependencies. A lack of continuous support for tools with different dependencies, as well as frequent and non-periodical versioning of the latter, often results in broken links and further compromises the reproducibility of analyses [36]. Widely used containerization technologies—e.g., Docker [28] and Singularity [29]—ensure reproducibility of software and replication of the analysis, thus partially addressing these challenges. By encapsulating software code along with all its corresponding dependencies in such containers, software packages become reproducible in any operating system in an easy-to-download-and-install fashion, on any infrastructure.

6.1.3 Contribution

The Institute of Marine Biology Biotechnology and Aqua-culture (IMBBC) has been developing a computing hub that, in conjunction with national and European Research Infrastructures (RIs), can support state-of-the-art marine research. The regional IMBBC HPC facility allows processing of data that derive from the Institute’s sequencing platforms and expeditions and from multiple external sources in the context of interdisciplinary studies. Here, we present insights from a thorough analysis of the research supported by the facility and some of its latest usage statistics in terms of resource requirements, computational methods, and data types; the above have contributed in shaping the facility along its lifespan.

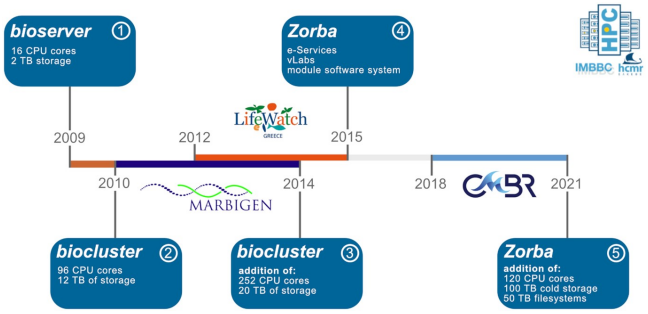

FIGURE 6.1: Evolution of the IMBBC HPC facility during the past 12 years, with hardware upgrades (blue boxes) and funding milestones (logos of RIs) highlighted. A single server that launched the bioinformatics era in 2009 evolved to the current Tier 2 system Zorba (Box 4), which allows processing of a wide variety of information from DNA sequences to biodiversity data. Different names of the facility denote distinct system architectures.

6.1.4 Methods

The IMBBC HPC Facility: From a Single Server to a Tier 2 System

The IMBBC HPC facility was launched in 2009 to support computational needs over a range of scientific fields in marine biology, with a focus on non-model taxa [39]. The facility was initiated as an infrastructure of the Institute of Marine Biology and Genetics of the Hellenic Centre for Marine Research. Its development has followed the development of national RIs (Fig. 1; also see Section A1 in Zafeiropoulos et al. [263]). The first nodes were used to support the analysis of data sets generated from methods such as eDNA metabarcoding and multiple omics. Since 2015, the facility also supports Virtual Research Environments, including e-services and virtual laboratories. The current configuration of the facility presented herein is named *Zorba* (Fig. 1, Box 4) and will be upgraded within 2021 (see Section Future Directions). Hereafter, *Zorba* refers to the specific system setup from 2015 and onwards, while the facility throughout its lifespan will be referred to as "IMBBC HPC".

Zorba currently consists of 328 CPU cores, 2.3 TB total memory, and 105 TB storage. Job submission takes place on the 4 available computing partitions, or queues, as explained in Fig. 2. *Zorba* at its current state achieves a peak performance of 8.3 trillion double-precision floating-point operations per second, or 8.3 Tflops, as estimated by LinPack benchmarking [264]. On top of these, a total 7.5 TB is distributed to all servers for the storage of environment and system files. Interconnection of both the compute and login nodes takes place via an infiniband interface with a capacity of 40 Gbps, which features very high throughput and very low latency. Infiniband is also used for a switched interconnection between the servers and the 4 available file systems. A thorough technical description of *Zorba* is available in Section A2 of Zafeiropoulos et al. [263].

More than 200 software packages are currently installed and available to users at *Zorba*, covering the most common analysis types. These tools allow assembly, HTS data preprocessing, phylogenetic tree construction, ortholog finding, and population structure

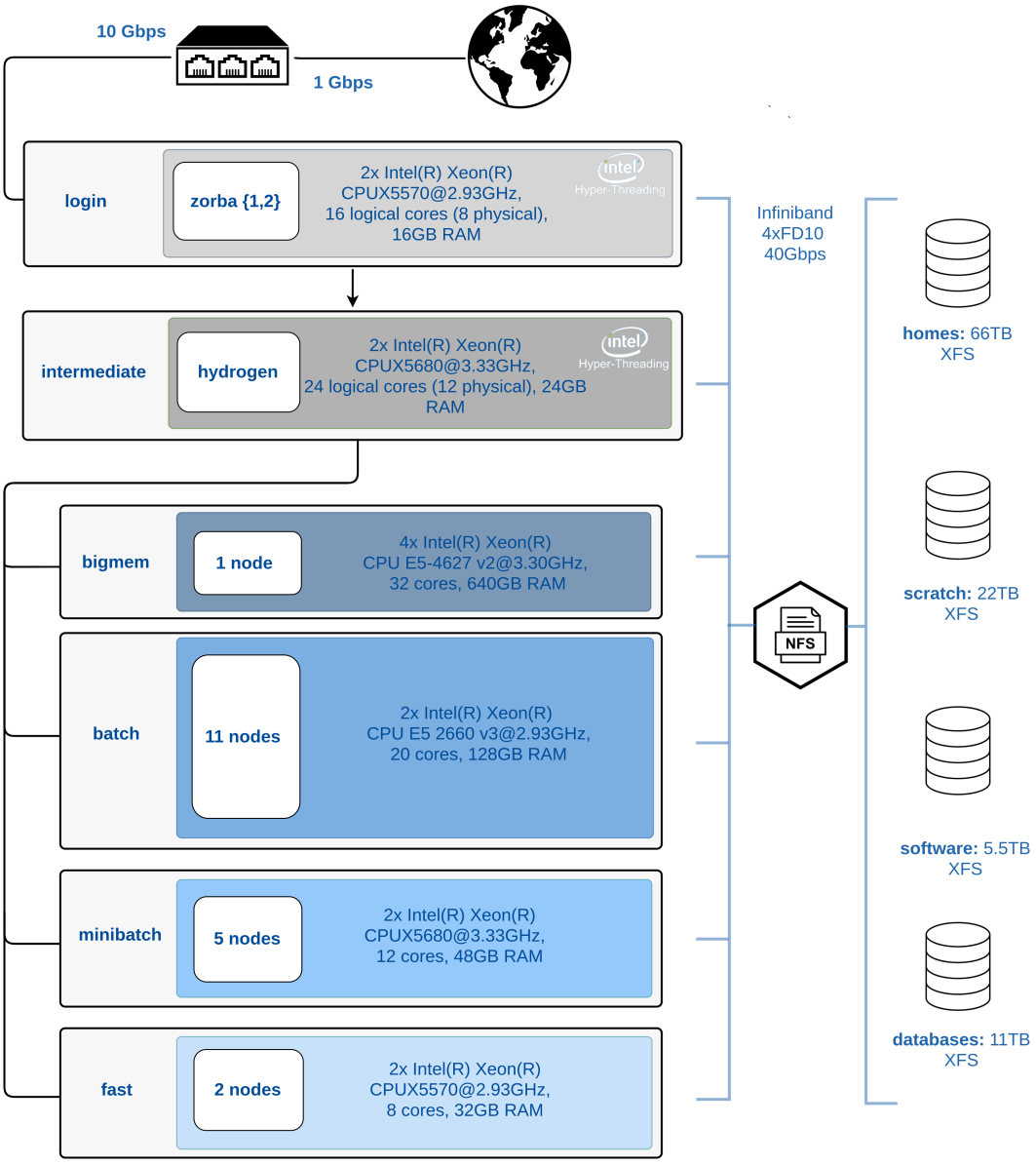

FIGURE 6.2: Block diagram of the *Zorba* architecture. This is the IMBBC HPC facility architecture in its current setup, after 12 years of development. There are 2 login nodes and 1 intermediate where users may develop their analyses. Computational nodes are split into 4 partitions with different specs and policy terms: bigmem supporting processes requiring up to 640 GB RAM, batch handling mostly (but not exclusively) parallel-driven jobs (either in a single node or across several nodes), minibatch aiming to serve parallel jobs with reduced resource requirements, and fast partition for non-intensive jobs. All servers, except file systems, run Debian 9 (kernel 4.9.0-8-amd64). CCBY icons from the Noun Project: "nfs file document icon" by IYIKON, PK; "Earth" By mungang kim, KR; "database": By Vectorstall, PK; "switch" by Bonegolem, IT

modeling, to name a few. Access to these packages is provided through Environment Modules, a broadly used means of accessing software in HPC systems [265].

During the last 2 years, *Zorba* has been moving from system-dependent pipelines previously developed at IMBBC (e.g., ParaMetabarCoding) towards containerization of available and new pipelines/tools. A complete metabarcoding analysis tool for various marker genes (PEMA) [101], the chained and automated use of STACKS, software for population genetics analysis from short-length sequences [266] (latest version), a set of statistical functions in R for the computation of biodiversity indices, and analyses in cases of high computational demands [267], as well as a programming workflow for the automation of biodiversity historical data curation (DECO) are among the in-house developed containers. The standard container/image format used on *Zorba* is Singularity. Singularity images can be served by any *Zorba* partition; Docker images can run instantly as Singularity images. A thorough description of the software containers developed in *Zorba* can be found in Section D of Zafeiropoulos et al. [263].

Zorba's daily functioning is ensured by a core team of 4 full-time, experienced staff: a hardware officer, 2 system administrators, and a permanent researcher in biodiversity informatics and data science.

More than 70 users (internal and external scientists), investigators, postdoctoral researchers, technicians, and doctoral/postgraduate students have gained access to the HPC infrastructure thus far. Support is provided officially through a help desk ticketing system. An average of 31 requests/month have been received (since June 2019), with the most demanded categories being troubleshooting (38.2%) and software installation (23.8%). Since October 2017, monthly meetings among HPC users have been established to regularly discuss such issues.

Proper scheduling of the submitted jobs and fair resource sharing is a major task that needs to be confronted day to day. To address this, a specific usage policy for each of the various partitions and a scheduling software tool set have been adopted in *Zorba*. Policy terms are dynamically adapted to the HPC hardware architecture and to the usage statistics, with revisions being discussed between the HPC core team and users. The Simple Linux Utility for Resource Management (SLURM) open-source cluster management system orchestrates the job scheduling and allocates resources, and a booking system helps users to organize their projects and administrators to monitor the resource reservations on a mid- to long-term basis. A SLURM Database Daemon (slurmdbd) has also been installed to allow logging and recording of job usage statistics into a separate SQL database (see Section C1 in Zafeiropoulos et al. [263]). An extended description of user and job administrations and orchestration can be found in Section C1 of Zafeiropoulos et al. [263]).

Training has been an integral component of the HPC facility mindset since its launch and enables knowledge sharing across MSc and PhD students and researchers within and outside the Institute. Introductory courses are organized on a regular basis, aimed at familiarizing new users with Unix environments, programming, and HPC usage policy and resource allocation (e.g., job submission in SLURM). Furthermore, the IMBBC HPC facility has served, since 2011, as an international training platform for specific types of bioinformatic analyses (see Section C2 in Zafeiropoulos et al. [263]). For instance, the facility has provided computational resources for workshops on Microbial Diversity,

Genomics and Metagenomics, Genomics in Biodiversity, Next-Generation Sequencing technologies and informatics tools for studying marine biodiversity and adaptation in the long term, or Ecological Data Analysis using R. The plan is to enhance and diversify the educational component of the HPC facility by providing courses on a more permanent basis and targeting a larger audience. An extensive listing of training activities is given in Section C2 of Zafeiropoulos et al. [263].

6.1.5 Results

Computational Breakdown of the IMBBC HPC-Supported Research

Systematic labelling of IMBBC HPC-supported published studies ($n = 47$) was performed to highlight their resource requirements. Each study was manually labelled with the relevant scientific field, the data acquisition method, the computational methods, and its resource requirements; all the annotations were validated by the corresponding authors (see Section D2 in Zafeiropoulos et al. [263]). It should be stated that the conclusions of this overview are specific to the studies conducted at IMBBC.

The scientific fields of Aquaculture (40% of studies), Biodiversity (26% of studies), and Organismal biology (19% of studies) account for the majority of the research publications supported by the IMBBC HPC facility (Fig. 3; Supplementary File imbbc_hpc_labelling_data.xlsx in Zafeiropoulos et al. [263]).

In comparison, studies in the Biotechnology and Agriculture fields indicate contemporary and beyond-marine orientations of research at IMBBC, respectively (see Section B2 in Zafeiropoulos et al. [40]). In addition, 8 methods of data acquisition (experimental or *in silico*) have been defined (Fig. 3). Among these methods, whole-genome sequencing and whole-transcriptome sequencing have been widely used in multiple fields (Biotechnology, Organismal Biology, Aquaculture). Conversely, Double digest restriction-site associated sequencing (ddRADseq) has been solely employed for population genetic studies in the context of Aquaculture.

The 47 published studies employed different computational methods (sets of tasks executed on the HPC facility). These studies served different purposes, from a range of bioinformatics analyses to HPC-oriented software optimization. The computational methods were categorized in 8 classes (Fig. 4). The resource requirements of each computational method were evaluated in terms of memory usage, computational time, and storage. Reflecting the current *Zorba* capacity, studies which, in any part of their analysis, exceeded 128 GB of memory or/and 48 hours of running time or/and 200 GB physical space were classified as studies with high demands (see Supplementary file imbbc_hpc_labelling_data.xlsx in [263]).

As shown in Fig. 4, the 2 most commonly used computational methods have rather different resource requirements. While DE analysis shows a notable trend for both long computational times (Fig. 4a) and high memory (Fig. 4b), eDNA-based community analysis does not have high resource requirements either in computation time or memory. High memory was commonly associated with computational methods, including de novo assembly; all relevant research concerned non-model taxa and involved short-read sequencing or combinations of short- and long-read sequencing. By contrast, phylogenetic

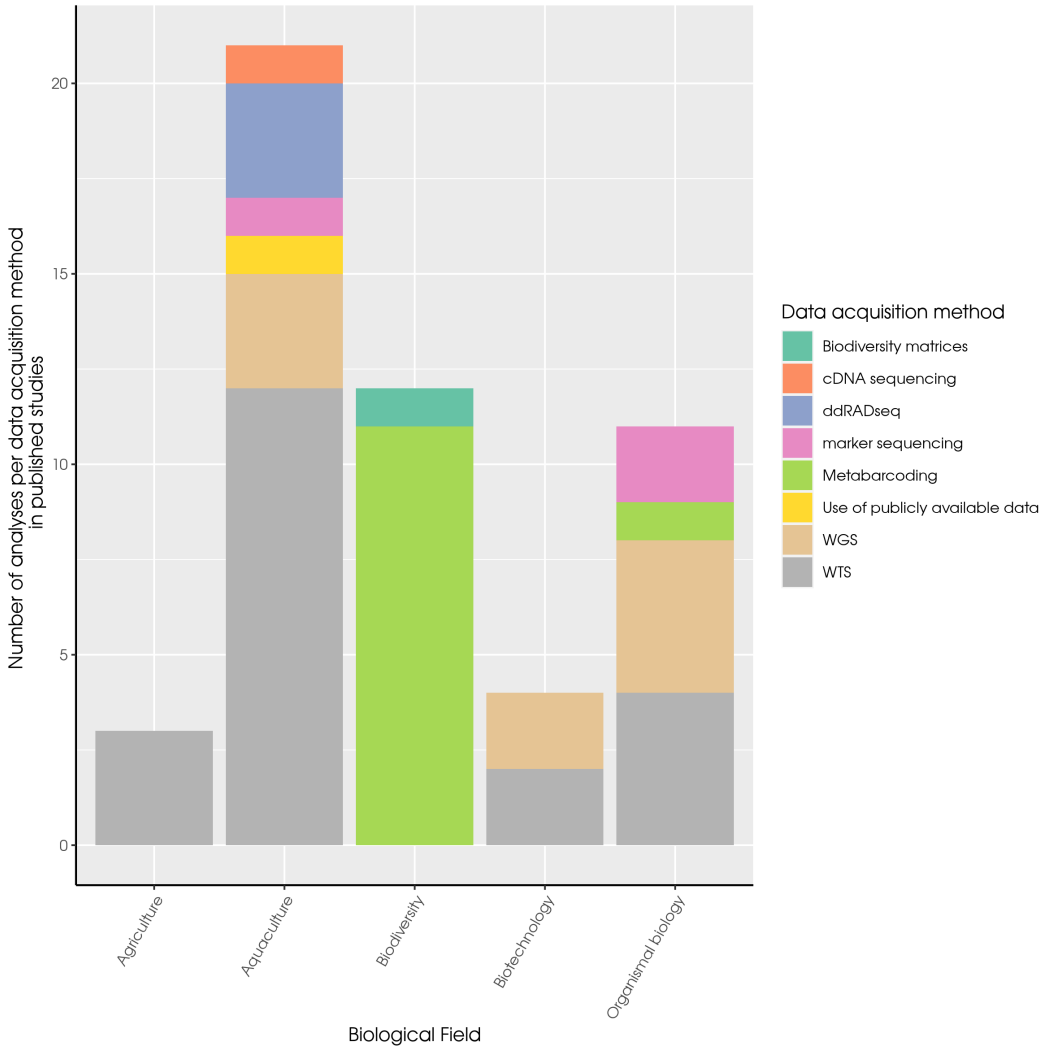

FIGURE 6.3: Bar chart with the number of publications that have used IMBBC HPC facility resources, grouped by scientific field. The different methods for data acquisition are also presented. WGS, whole-genome sequencing; WTS, whole-transcriptome sequencing.

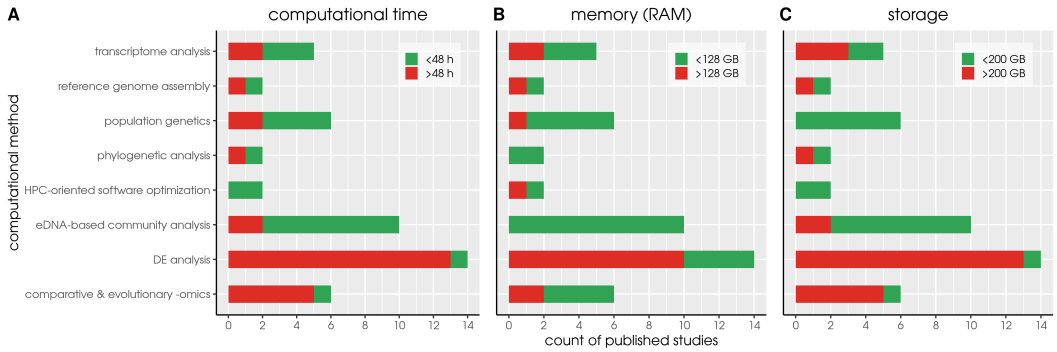

FIGURE 6.4: Red bars denote published research with high resource requirements of the various computational methods employed at the IMBBC HPC facility due to (a) long computational times (> 48 h), (b) high memory requirements (> 128 GB), or (c) high storage requirements (> 200 GB). For instance, no eDNA-based community analyses performed at *Zorba* thus far have required a large amounts of memory.

analysis studies did not involve intensive RAM use; this is largely due to the fact that software used by IMBBC users adopts parallel solutions for tree construction. Long computational times (Fig. 4a) were most often observed at the functional annotation step in transcriptome analysis, DE analysis, and comparative and evolutionary omics, when this step involved BLAST queries of thousands of predicted genes against large databases, such as nr (NCBI). Finally, a common challenge emerging from all bioinformatic approaches is significant storage limitations (Fig. 4c); this challenge was associated with the use of HTS technologies that produce large amounts of raw data, the analysis of which involves the creation of numerous intermediate files.

Overall, published studies using the IMBBC HPC facility show a degree of variance with respect to the types of tools used (depending on the user, their bioinformatic literacy, and other factors), each of which is more or less optimized with respect to HPC use. Moreover, the variance in computational needs observed within each type of computational method reflects the diversity of the studied taxonomic groups. For instance, transcriptome analysis (involving de novo assembly and functional annotation steps) was employed for the study of taxa as diverse as bacteria, sponges, fungi, fish, and goose barnacles. The complexity of each of these organisms' transcriptomes can, to a large extent, explain the differences observed in computational time, memory, and storage.

Furthermore, *Zorba* CPU and RAM statistics collected since 2019 displayed some overall patterns, including an average computation load per month of less than or close to 50% of its max capacity (50% of 236 kilocorehours/month) for most (20) of the 24 months of the logging period. Memory requirements were also heterogeneous: most (90%) of the 44,000 jobs performed in the same 24-month period required less than 10 GB of RAM, and 0.30% of the jobs required more than 128 GB of RAM (i.e., exceeding the memory capacity of the main compute nodes [batch partition]). The detailed usage statistics of *Zorba* are described in Section B1 and Supplementary file zorba_usage_statistics.xlsx of Zafeiropoulos et al. [263].

6.1.6 Discussion

Scientific Impact Stories

Below, some examples of research results that were made possible with the IMBBC HPC facility are described. This list of use cases is by no means exhaustive, but rather an attempt to highlight different fields of research supported by the facility, along with their distinct computational features.

Invasive species range expansion detected with eDNA data from Autonomous Reef Monitoring Structures

The Mediterranean biodiversity and ecosystems are experiencing profound transformations owing to Lessepsian migration, international shipping, and aquaculture, which lead to the migration of nearly 1,000 alien species [46]. The first step towards addressing the effects of these invasions is monitoring of the introduced taxa. A powerful tool in this direction has been eDNA metabarcoding, which has enchanced detection of invasive species [268], often preceding macroscopic detection. One such example is the first record of the nudibranch *Anteaeolidiella lurana* (Ev. Marcus & Er. Marcus, 1967) in Greek waters in 2020 [269]. An eDNA metabarcoding analysis allowed for detection of the species with high confidence on fouling communities developed on Autonomous Reef Monitoring Structures (ARMS). This finding, confirmed with image analysis of photographic records on a later deployment period, is an example of work conducted within the framework of the European ASSEMBLE plus programme ARMS-MBON (Marine Biodiversity Observation Network). PEMA software [101] was used in this study, as well as in the 30-month pilot phase of ARMS-MBON [270].

Providing omics resources for large genome-size, non-model taxa

Zorba has been used for building and annotating numerous de novo genome and transcriptome assemblies of marine species, such as the gilthead sea bream *Sparus aurata* [271] or the greater amberjack *Seriola dumerili* [272]. Both genome and transcriptome assemblies of species with large genomes often exceed the maximum available memory limit, eventually affecting the strategic choices for *Zorba* future upgrades (see Section Future Directions). For instance, building the draft genome assembly of the seagrass Halophila stipulacea (estimated genome size 3.5 GB) using Illumina short reads has been challenging even for seemingly simple tasks, such as a kmer analysis [273]. Taking advantage of short- and long-read sequencing technologies to construct high-quality reference genomes, the near-chromosome level genome assembly of Lagocephalus sceleratus (Gmelin, 1789) was recently completed as a case study of high ecological interest due to the species' successful invasion throughout the Eastern Mediterranean [274]. In the context of this study, an automated containerized pipeline allowing high-quality genome assemblies from Oxford Nanopore and Illumina data was developed (SnakeCube [275]). The availability of standardized pipelines offers great perspective for in-depth studies of numerous marine species of interest in aquaculture and conservation biology, including

rigorous phylogenomic analyses to position each species in the tree of life (e.g., Natsidis et al. [276]).

DE analysis of aquaculture fish species sheds light on critical phenotypes

Distinct, observable properties, such as morphology, development, and behavior, characterize living taxa. The corresponding phenotypes may be controlled by the interplay between specific genotypes and the environment. To capture an individual's genotype at a specific time point, molecular tools for transcript quantification have followed the fast development of technologies, with Expressed Sequence Tags as the first approach to be historically used, especially suited for non-model taxa [56]. Nowadays, the physiological state of aquaculture species is retrieved through investigation of stage-specific and immune- and stress response–specific transcriptomic profiles using RNAseq. The corresponding computational workflows involve installing various tools at *Zorba* and implementing a series of steps that often take days to compute. These analyses, besides detecting transcripts at a specific physiological state, have successfully identified regulatory elements, such as microRNAs. Through the construction of a regulatory network with putative target genes, microRNAs have been linked to the transcriptome expression patterns. The most recent example is the identification of microRNAs and their putative target genes involved in ovary maturation [277].

Large-scale ecological statistics: are all taxa equal?

The nomenclature of living organisms, as well as their descriptions and their classifications under a specific nomenclature code, have been studied for more than 2 centuries. Up to now, all the species present in an ecosystem have been considered equal in terms of their contributions to diversity. However, this axiom has been tested only once before, on the United Kingdom's marine animal phyla, showing the inconsistency of the traditional Linnaean classification between different major groups [278]. In Arvanitidis et al. [279], the average taxonomic distinctness index ($\Delta+$) and its variation ($Lambda+$) were calculated on a matrix deriving from the complete World Register of Marine Species [280], containing more than 250,000 described species of marine animals. It is the R-vLab web application, along with its HPC high RAM back-end components (on bigmem, see Section The IMBBC HPC Facility: From a Single Server to a Tier 2 System) that made such a calculation possible. This is the first time such a hypothesis has been tested on a global scale. Preliminary results show that the 2 biodiversity indices exhibit complementary patterns and that there is a highly significant yet non-linear relationship between the number of species within a phylum and the average distance through the taxonomic hierarchy.

Discovery of novel enzymes for bioremediation

Polychlorinated biphenyls are complex, recalcitrant pollutants that pose a serious threat to wildlife and human health. The identification of novel enzymes that can degrade such organic pollutants is being intensively studied in the emerging field of bioremediation. In the context of the Horizon 2020 Tools And Strategies to access original bioactive

compounds by Cultivating MARine invertebrates and associated symbionts (TASCMAR) project, global ocean sampling provided a large biobank of fungal invertebrate symbionts and, through large-scale screening and bioreactor culturing, a marine-derived fungus able to remove a polychlorinated biphenyl compound was identified for the first time. *Zorba* resources and domain expertise in fungal genomics were used as a Centre for the Study and Sustainable Exploitation of Marine Biological Resources (CMBR) service for the analysis of multi-omic data for this symbiont. Following genome assembly of *Cladosporium sp.* TM-S3 [281], transcriptome assembly and a phylogenetic analysis revealed the full diversity of the symbiont's multicopper oxidases, enzymes commonly involved in oxidative degradation [282]. Among these, 2 laccase-like proteins shown to remove up to 71% of the polychlorinated biphenyl compound are now being expressed to optimize their use as novel biocatalysts. This step would not have been possible without the annotation of the Cladosporium genome with transcriptome data; mapping of the purified enzymes' LC-MS (Liquid chromatography–mass spectrometry) spectra against the set of predicted proteins allowed for identification of their corresponding sequences.

Lessons Learned

Depth and breadth are both required for a bioinformatics-oriented HPC

In our experience, the vast majority of the analyses run at the IMBBC HPC infrastructure are CPU-intensive. RAM-intensive jobs (> 128 GB RAM, see Section Computational Breakdown of the IMBBC HPC-Supported Research) represent only 0.3% of the total jobs executed over the last 2 years (see Section B1 in [263]). Despite the difference in the frequency of executed jobs with distinct requirements, serving both types of jobs and ensuring their successful completion is equally important for addressing fundamental marine research questions (as shown in Section Computational Breakdown of the IMBBC HPC-Supported Research). The need for both HPC depth (a few high-memory nodes) and breadth (a number of slimmer nodes) has been previously reported [258]. This need reflects the idiosyncrasy of different bioinformatics analysis steps, often even within the same workflow. High-memory nodes are necessary for tasks such as de novo assembly of large genomes, while the availability of as many less powerful nodes as possible can speed up the execution of less demanding tasks and free resources for other users. Future research directions and the available budget further dictate tailoring of the HPC depth and breadth. Cloud-based services—e.g., for containerized workflows—may also facilitate this process once these become more affordable.

Quota . . . overloaded

We observed that independently of the type of analysis, storage was an issue for all *Zorba* users (Fig. 4). A high percentage of these issues relate to the raw data from HTS projects. These data are permanently stored in the home directories, occupying significant space. This, in conjunction with the fact that users delete their data with great reluctance, makes storage a major issue of daily use in *Zorba*. In specific cases where users' quota was exceeded uncontrollably, the *Zorba* team has been applying compression of raw and output data in contact with the user, but this is by no means a stable strategy. More

generally, with the performance of the existing storage configuration in *Zorba* close to reaching its limits due to the increase in users and its concurrent use, several solutions have been adopted to resolve the issue. The most long-lasting solution has been the adoption of a per user quota system to allow storage sustainability and fairness in our allocation policy. This quota system nevertheless constitutes a limiting factor in pipeline execution, since lots of software tools produce unpredictably too many intermediate files, which not only increase storage but also cause job failures due to space restrictions. We managed the above issue by adding a scratch file system as an intermediate storage area for the runtime capacity needs. Following completion of their analysis, a user retains only the useful files and the rest are permanently removed. A storage upgrade scheduled within 2021 (see Section Future Directions) is expected to alleviate current storage challenges in *Zorba*. However, given the ever-increasing data production (e.g., as the result of decreasing sequencing costs and/or of rising imaging technologies), the responsible storage use approaches described here remain only partial solutions to anticipated future storage needs. Centralized (Tier 1 or higher) storage solutions represent a longer-term solution, which is in line with current views on how to handle big data generated by international research consortia in a long-lasting manner.

Continuous intercommunication among different disciplines matters

Smooth functioning of an HPC system and exploitation of its full potential for research requires stable employment of a core team of computer scientists and engineers, in close collaboration with an extended team of researchers. At least 4 disciplines are involved in *Zorba*-related issues: computer scientists, engineers, biologists (in the broad sense, including ecologists, genomicists, etc.), and bioinformaticians with varying degrees of literacy in biology and informatics and various domain specializations (comparative genomics, biodiversity informatics, bacterial metagenomics, etc). The continuous communication among representatives of these 4 disciplines has substantially contributed to research supported by *Zorba* and to the evolution of the HPC system itself over time. In our experience, an HPC system cannot function effectively and for long without full-time system administrators, nor with bioinformaticians alone. Although it has not been the case since the system's onset, investment in monthly meetings, seminars, and training events (in biology, containers, domain-specific applications, and computer science; see Section The IMBBC HPC Facility: From a Single Server to a Tier 2 System) is the only way to establish stable intercommunication among different players of an HPC system. Such proximity translates into timely and adequate systems and bioinformatics analysis support, an element that in its turn translates into successful research (see Section Computational Breakdown of the IMBBC HPC-Supported Research). It should be noted that the overall good experience in connectivity among different HPC players derives from *Zorba* being a Tier 2 system, with a number of active permanent users in double digits. The establishment of such inter-communication was relatively straightforward to implement with periodic meetings and the assistance of ticketing and other management solutions (see Section C1 in [263]).

The way forward: develop locally and share and deploy centrally

The various approaches regarding the function of an HPC system are strongly related to the different viewpoints of the academic communities towards the relatively new disciplines of bioinformatics and big data. These approaches are strongly affected by national and international decisions that affect the ability to fund supercomputer systems. There are advantages in deploying bioinformatics-oriented HPC systems in centralized (Tier 0 and Tier 1) facilities: better prices at hardware purchases, easier access to HPC-tailored facilities (for instance, in terms of the cooling system and physical space), or experienced technical personnel (see also [258]). However, synergies between regional (Tier 2) and centralized HPC systems are fundamental for moving forward in supporting the diverse and demanding needs of bioinformatics. An example of such synergies concerns technical solutions (e.g., containerization) that address long-standing software sharing issues. In our experience, a workflow/pipeline can be developed by experts within the context of a specific project in a regional HPC facility. Once a production version of the pipeline is packaged, it can be distributed to centralized systems to cover a broader user audience (see Section The IMBBC HPC Facility From a Single Server to a Tier 2 System). Singularity containers have been developed to utterly suit HPC environments, mostly because they permit root access of the system in all cases. In addition, Singularity is compatible with all Docker images and can be used with Graphics Processing Units (GPUs) and Message Passing Interface (MPI) applications. This is why we chose to run containers in a Singularity format at *Zorba*. However, as Docker containers are widely used, especially in cloud computing (see more about cloud computing in Section Cloud Computing), workflows and services produced at IMBBC are offered in both container formats. Containers are an already established technology, used by the biggest cloud providers worldwide and increasingly by non-profit research institutes. Despite indirect costs (e.g., costs to containerize legacy software), we believe that these technologies will become the norm in the future, especially in the context of reproducibility and interoperability of bioinformatics analysis.

Software optimizations for parallel execution

The most common ways of achieving implicit or explicit parallelization in modern multicore systems for bioinformatics, computational biology, and systems biology software tools are the software threads—provided by programming languages—and/or the OpenMP API [283]. These types of multiprocessing make good use of the available cores on a multicore system (single node), but they are not capable of combining the available CPU cores from more than 1 node. Some other software tools use MPI to spawn processing chunks to many servers and/or cores or (even better) combine MPI with OpenMP/Threads to maximize the parallelization in hybrid models of concurrency. Such designs are now used to a great extent in some cases, such as phylogeny inference software that makes use of Monte Carlo Markov Chain samplers. However, these cases are but a small number compared to the majority of bioinformatics tasks, while their usage in other analyses is low. At the hardware level, simultaneous multithreading is not enabled in the compute nodes of the IMBBC HPC infrastructure. Since the majority of analyses running

on the cluster demand dedicated cores, hardware multithreading does not perform well. In our experience, the existence of more (logical) cores in compute nodes misleads the least experienced users into using more threads than the physically available ones, which slows down their executions. In comparison, assisting servers (filesystems, login nodes, web servers) make use of hardware multithreading, since they serve numerous small tasks from different users/sources that commonly contain Input/Output (I/O) operations. GPUs provide an alternative way for parallel execution, but they are supported by a limited number of bioinformatics software tools. Nevertheless, GPUs can optimize the execution process in specific, widely used bioinformatic analyses, such as sequence alignment [284, 285], image processing in microtomography (e.g., microCT), or basecalling of Nanopore raw data.

Cloud Computing

A recent alternative to traditional HPC systems, such as that described in this review, is cloud computing. Cloud computing is the way of organizing computing resources so they are provided over the Internet ("the cloud"). This paradigm of computing requires the minimum management effort possible [286]. Cloud computing providers exist in both commercial vendors and academic/publicly funded institutions and infrastructures (for more on cloud computing for bioinformatics, see Langmead and Nellore [287]). Computing resources can be reserved from individuals, institutions, organizations, or even scientific communities. The most widely-known commercial cloud providers are the "big 3" of cloud computing—namely, Amazon Web Services, Google Cloud Platform, and Microsoft Azure — while other cloud vendors are constantly emerging. Academic/publicly funded providers are also available: e.g., the EMBL–EBI Embassy Cloud.

Cloud computing services are being increasingly adopted in research, mainly because they offer simplicity and high availability to users with reduced or even no experience in HPC systems, through web interfaces. For this type of user, the time needed for data manipulation, software installation, and user-system interaction is significantly reduced compared to using a local HPC facility.

Container technologies, especially Docker, along with container-management systems such as Kubernetes combined with OpenStack, have been widely used in a number cloud computing systems, in particular in the research domain. It should be noted, however, that tool experimentation and benchmarking is more limited in cloud computing compared to local facilities and is costly, since it demands additional core hours of segmented computation. In-house HPC infrastructures can be fully configured to suit specific research area needs (storage available, fast interconnection for MPI jobs, number of CPUs versus available RAM, assisting services, etc.). Moreover, in cases where InfiniBand interconnection, a computer networking communications standard, is adopted in HPC, the performance in jobs and software that take advantage of it is substantial. Given the features and advantages of each approach (mentioned above) one could foresee the scenario of combining them to address the research community needs.

Future Directions

An upgrade of the existing hardware design of Zorba has been scheduled in 2021, funded by the CMBR research infrastructure (Fig. 1). More specifically:

3 nodes of 40 CPU physical cores will be added through new partitions (120 cores in total); the total RAM will be increased by 3.5 TB; 100 TB of cold storage will be installed and is expected to alleviate the archiving problem at the existing homes/scratch file systems; and the total usable existing storage capacity for users in home and scratch partitions will be increased by approximately 100 TB.

With this upgrade, it is expected that the total computational power of *Zorba* will be increased by approximately 6 TFlops, while the infrastructure will be capable of serving memory-intensive jobs requiring up to 1.5 TB of RAM, hosted on a single node. Eventually, more users will be able to concurrently load and analyze big data sets on the file systems. Over the coming 2 years, *Zorba* is also expected to have 2 major additions:

- the acquisition of a number of GPU nodes to build a new partition, especially for serving software that has been ported to run on GPUs; and
- the design of a parallel file system (Ceph or Lustre) to optimize concurrent I/O operations to speed up CPU-intensive jobs.

The expectation is that the upcoming upgrade of *Zorba* will further enhance collaborations with external users, since the types of bioinformatic tasks supported by the infrastructure are common to other disciplines beyond marine science, such as environmental omics research in the broad term. A nationwide survey targeting the community of researchers studying the environment and adopting the same approaches (HTS, biodiversity monitoring) has revealed that their computational and training needs are on the rise (A. Gioti et al., unpublished observations). Usage peaks and valleys were observed in *Zorba* (see Section B1 in [263]), similarly to other HTS-oriented HPC systems [258]. It is therefore feasible to share *Zorba*'s idling time with other scientific communities. Besides, the *Zorba* upgrade is very timely in coming during a period where additional computational infrastructures emerge: the Cloud infrastructure Hypatia, funded by the Greek node of ELIXIR, is entering its production phase. It will constitute a national Tier 1 HPC facility, designed to host 50 computational nodes of different capabilities (regular servers, GPU-enabled servers, Solid-State Drive-enabled servers, etc.) and provide users the option to either create custom virtual machines for their computational services or to upload and execute workflows of containerized scientific software packages. In this context, a strategic combination of *Zorba* and Hypatia is expected to contribute to a strong computational basis in Greece. It is also expected that *Zorba* functionality will be augmented also through its connection with the Super Computing Installations of LifeWatch ERIC (European Research Infrastructure Consortium) (e.g., Picasso facility in Malaga, Spain). Building upon the lessons learned in the last 12 years, a foreseeable challenge for the facility is the enhancement of its usage monitoring to the example of international HPC systems [288], in order to allow even more efficient use of computational resources.

6.1.7 Conclusions

Zorba is an established Tier 2 HPC regional facility operating in Crete, Greece. It serves as an interdisciplinary computing hub in the eastern Mediterranean, where studies in marine conservation, invasive species, extreme environments, and aquaculture are of great scientific and socio-economic interest. The facility has supported, since its launch over a decade ago, a number of different fields of marine research, covering all kingdoms of life; it can also share part of its resources to support research beyond the marine sciences.

The operational structure of *Zorba* enables continuous communication between users and administrators for more effective user support, troubleshooting, and job scheduling. More specifically, training, regular meetings, and containerization of in-house pipelines have proven constructive for all teams, students, and collaborators of IMBBC. This operational structure has evolved over the years based on the needs of the facility's users and the available resources. The practical solutions adopted—from hardware (e.g., depth/breadth balanced structure, user quotas, and temporary storage) to software (e.g., modularized bioinformatics application maintenance and containerization) and human resource management (e.g., frequent intercommunication, continuous cross-discipline training)—reflect IMBBC research to a large extent. However, and by incrementing previous reviews [258], other Institutes and HPC facilities can be informed on the lessons learned (see Section Lessons Learned), and reflect on the computational requirement analysis of the methods presented (see Section Computational Breakdown of the IMBBC HPC-Supported Research) through the spectrum of their own research so as to plan ahead.

HPC facilities could reach a benefit greater than the sum of their capacities once they interconnect. The IMBBC HPC facility lies at the crossroad of 3 RIs, CMBR (Greek node of EMBRC-ERIC), LifeWatchGreece (Greek node of LifeWatch ERIC), and ELIXIR Greece, and via these will pursue further collaboration at larger Tier 0 and Tier 1 levels.

Chapter 7

Conclusions

1. Role of technologies such as containerization.
2. Trends for reproducible pipelines and role of infrastuctures

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).

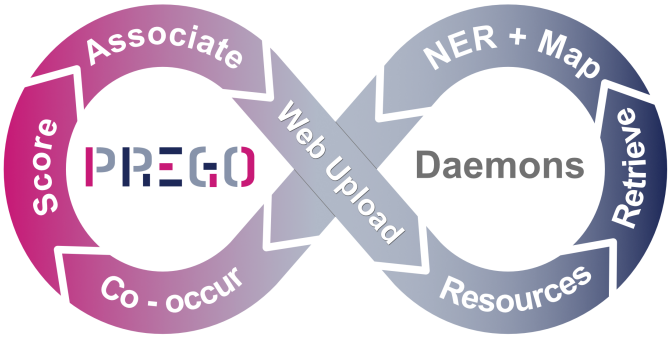

FIGURE A.1: Software daemons perform all steps of the PREGO methodology in a continuous manner similar to the Continuous Development and Continuous Integration method.

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 [143] and COMPARTMENTS [289], 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:

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

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 [289]. 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 **??**. 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] N. Swainston, K. Smallbone, H. Hefzi, P. D. Dobson, J. Brewer, M. Hanscho, D. C. Zielinski, K. S. Ang, N. J. Gardiner, J. M. Gutierrez, S. Kyriakopoulos, M. Lakshmanan, S. Li, J. K. Liu, V. S. Martínez, C. A. Orellana, L.-E. Quek, A. Thomas, J. Zanghellini, N. Borth, D.-Y. Lee, L. K. Nielsen, D. B. Kell, N. E. Lewis, and P. Mendes, "Recon 2.2: from reconstruction to model of human metabolism," *Metabolomics*, vol. 12, p. 109, June 2016.

[2] E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Dräger, N. Mih, F. Gatto, A. Nilsson, G. A. P. Gonzalez, M. K. Aurich, *et al.*, "Recon3D enables a three-dimensional view of gene variation in human metabolism," *Nature biotechnology*, vol. 36, no. 3, p. 272, 2018.

[3] J. Belilla, D. Moreira, L. Jardillier, G. Reboul, K. Benzerara, J. M. López-García, P. Bertolino, A. I. López-Archilla, and P. López-García, "Hyperdiverse archaea near life limits at the polyextreme geothermal dallol area," *Nature ecology & evolution*, vol. 3, no. 11, pp. 1552–1561, 2019.

[4] P. G. Falkowski, T. Fenchel, and E. F. Delong, "The microbial engines that drive earth's biogeochemical cycles," *science*, vol. 320, no. 5879, pp. 1034–1039, 2008.

[5] Y. M. Bar-On, R. Phillips, and R. Milo, "The biomass distribution on earth," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. 6506–6511, 2018.

[6] H. A. Rees, A. C. Komor, W.-H. Yeh, J. Caetano-Lopes, M. Warman, A. S. Edge, and D. R. Liu, "Improving the dna specificity and applicability of base editing through protein engineering and protein delivery," *Nature communications*, vol. 8, no. 1, pp. 1–10, 2017.

[7] L. Röttjers and K. Faust, "From hairballs to hypotheses–biological insights from microbial networks," *FEMS microbiology reviews*, vol. 42, no. 6, pp. 761–780, 2018.

[8] R. Cavicchioli, W. J. Ripple, K. N. Timmis, F. Azam, L. R. Bakken, M. Baylis, M. J. Behrenfeld, A. Boetius, P. W. Boyd, A. T. Classen, *et al.*, "Scientists' warning to humanity: microorganisms and climate change," *Nature Reviews Microbiology*, vol. 17, no. 9, pp. 569–586, 2019.

[9] W. Commons, "File:sulfur cycle for hydrothermal vents.png — wikimedia commons, the free media repository," 2020. [Online; accessed 30-December-2021].

[10] W. Commons, "File:nitrogen cycle for hydrothermal vents.png — wikimedia commons, the free media repository," 2020. [Online; accessed 30-December-2021].

[11] K. Deiner, H. M. Bik, E. Mächler, M. Seymour, A. Lacoursière-Roussel, F. Altermatt, S. Creer, I. Bista, D. M. Lodge, N. De Vere, *et al.*, "Environmental dna metabarcoding: Transforming how we survey animal and plant communities," *Molecular ecology*, vol. 26, no. 21, pp. 5872–5895, 2017.

[12] K. M. Ruppert, R. J. Kline, and M. S. Rahman, "Past, present, and future perspectives of environmental dna (edna) metabarcoding: A systematic review in methods, monitoring, and applications of global edna," *Global Ecology and Conservation*, vol. 17, p. e00547, 2019.

[13] P. Taberlet, E. Coissac, M. Hajibabaei, and L. H. Rieseberg, "Environmental dna," 2012.

[14] M. Stat, M. J. Huggett, R. Bernasconi, J. D. DiBattista, T. E. Berry, S. J. Newman, E. S. Harvey, and M. Bunce, "Ecosystem biomonitoring with edna: metabarcoding across the tree of life in a tropical marine environment," *Scientific Reports*, vol. 7, no. 1, pp. 1–11, 2017.

[15] Y. Ji, L. Ashton, S. M. Pedley, D. P. Edwards, Y. Tang, A. Nakamura, R. Kitching, P. M. Dolman, P. Woodcock, F. A. Edwards, *et al.*, "Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding," *Ecology letters*, vol. 16, no. 10, pp. 1245–1257, 2013.

[16] B. E. Deagle, S. N. Jarman, E. Coissac, F. Pompanon, and P. Taberlet, "Dna metabarcoding and the cytochrome c oxidase subunit i marker: not a perfect match," *Biology letters*, vol. 10, no. 9, p. 20140562, 2014.

[17] P. Ten Hoopen, R. D. Finn, L. A. Bongo, E. Corre, B. Fosso, F. Meyer, A. Mitchell, E. Pelletier, G. Pesole, M. Santamaria, *et al.*, "The metagenomic data life-cycle: standards and best practices," *GigaScience*, vol. 6, no. 8, p. gix047, 2017.

[18] R. Strohman, "Maneuvering in the complex path from genotype to phenotype," *Science*, vol. 296, no. 5568, pp. 701–703, 2002.

[19] M. Polanyi, "Life's irreducible structure: Live mechanisms and information in dna are boundary conditions with a sequence of boundaries above them," *Science*, vol. 160, no. 3834, pp. 1308–1312, 1968.

[20] A. Pavan-Kumar, P. Gireesh-Babu, and W. Lakra, "Dna metabarcoding: a new approach for rapid biodiversity assessment," *J Cell Sci Mol Biol*, vol. 2, no. 1, p. 111, 2015.

[21] P. F. Thomsen and E. Willerslev, "Environmental dna–an emerging tool in conservation for monitoring past and present biodiversity," *Biological conservation*, vol. 183, pp. 4–18, 2015.

[22] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[23] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, *et al.*, "Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science," tech. rep., PeerJ Preprints, 2018.

[24] F. Hildebrand, R. Tadeo, A. Y. Voigt, P. Bork, and J. Raes, "Lotus: an efficient and user-friendly otu processing pipeline," *Microbiome*, vol. 2, no. 1, pp. 1–7, 2014.

[25] J. Axtner, A. Crampton-Platt, L. A. Hörig, A. Mohamed, C. C. Xu, D. W. Yu, and A. Wilting, "An efficient and robust laboratory workflow and tetrapod database for larger scale environmental dna studies," *GigaScience*, vol. 8, no. 4, p. giz029, 2019.

[26] H. S. Gweon, A. Oliver, J. Taylor, T. Booth, M. Gibbs, D. S. Read, R. I. Griffiths, and K. Schonrogge, "Pipits: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the i llumina sequencing platform," *Methods in ecology and evolution*, vol. 6, no. 8, pp. 973–980, 2015.

[27] P. Cingolani, R. Sladek, and M. Blanchette, "Bigdatascript: a scripting language for data pipelines," *Bioinformatics*, vol. 31, no. 1, pp. 10–16, 2015.

[28] B. B. Rad, H. J. Bhatti, and M. Ahmadi, "An introduction to docker and analysis of its performance," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 17, no. 3, p. 228, 2017.

[29] G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific containers for mobility of compute," *PloS one*, vol. 12, no. 5, p. e0177459, 2017.

[30] E. Coissac, T. Riaz, and N. Puillandre, "Bioinformatic challenges for dna metabarcoding of plants and animals," *Molecular ecology*, vol. 21, no. 8, pp. 1834–1847, 2012.

[31] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *The ISME journal*, vol. 11, no. 12, pp. 2639–2643, 2017.

[32] C. Pauvert, M. Buée, V. Laval, V. Edel-Hermann, L. Fauchery, A. Gautier, I. Lesur, J. Vallance, and C. Vacher, "Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline," *Fungal Ecology*, vol. 41, pp. 23–33, 2019.

[33] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, "Vsearch: a versatile open source tool for metagenomics," *PeerJ*, vol. 4, p. e2584, 2016.

[34] X. Hao, R. Jiang, and T. Chen, "Clustering 16s rrna for otu prediction: a method of unsupervised bayesian clustering," *Bioinformatics*, vol. 27, no. 5, pp. 611–618, 2011.

[35] F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn, "Swarm v2: highly-scalable and high-resolution amplicon clustering," *PeerJ*, vol. 3, p. e1420, 2015.

[36] A. Lanzén, S. L. Jørgensen, D. H. Huson, M. Gorfer, S. H. Grindhaug, I. Jonassen, L. Øvreås, and T. Urich, "Crest–classification resources for environmental sequence tags," *PloS one*, vol. 7, no. 11, p. e49334, 2012.

[37] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.

[38] M. C. Rillig, M. Ryo, A. Lehmann, C. A. Aguilar-Trigueros, S. Buchert, A. Wulf, A. Iwasaki, J. Roy, and G. Yang, "The role of multiple global change factors in driving soil functions and microbial biodiversity," *Science*, vol. 366, no. 6467, pp. 886–890, 2019.

[39] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, "Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference," *Bioinformatics*, vol. 35, no. 21, pp. 4453–4455, 2019.

[40] P. Barbera, A. M. Kozlov, L. Czech, B. Morel, D. Darriba, T. Flouri, and A. Stamatakis, "Epa-ng: massively parallel evolutionary placement of genetic sequences," *Systematic biology*, vol. 68, no. 2, pp. 365–369, 2019.

[41] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy," *Applied and environmental microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[42] R. J. Machida, M. Leray, S.-L. Ho, and N. Knowlton, "Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples," *Scientific data*, vol. 4, no. 1, pp. 1–7, 2017.

[43] P. J. McMurdie and S. Holmes, "phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data," *PloS one*, vol. 8, no. 4, p. e61217, 2013.

[44] "Fastqc," Jun 2015.

[45] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[46] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet. journal*, vol. 17, no. 1, pp. 10–12, 2011.

[47] S. I. Nikolenko, A. I. Korobeynikov, and M. A. Alekseyev, "Bayeshammer: Bayesian clustering for error correction in single-cell sequencing," in *BMC genomics*, vol. 14, pp. 1–11, Springer, 2013.

[48] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, *et al.*, "Spades: a new genome assembly algorithm and its applications to single-cell sequencing," *Journal of computational biology*, vol. 19, no. 5, pp. 455–477, 2012.

[49] A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld, "Pandaseq: paired-end assembler for illumina sequences," *BMC bioinformatics*, vol. 13, no. 1, pp. 1–7, 2012.

[50] F. Boyer, C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet, and E. Coissac, "obitools: A unix-inspired software package for dna metabarcoding," *Molecular ecology resources*, vol. 16, no. 1, pp. 176–182, 2016.

[51] R. C. Edgar, "Search and clustering orders of magnitude faster than blast," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[52] R. H. Nilsson, K.-H. Larsson, A. F. S. Taylor, J. Bengtsson-Palme, T. S. Jeppesen, D. Schigel, P. Kennedy, K. Picard, F. O. Glöckner, L. Tedersoo, *et al.*, "The unite database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications," *Nucleic acids research*, vol. 47, no. D1, pp. D259–D264, 2019.

[53] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt, and E. W. Sayers, "Genbank.," *Nucleic acids research*, vol. 46, no. D1, pp. D41–D47, 2018.

[54] L. Czech, P. Barbera, and A. Stamatakis, "Methods for automatic reference trees and multilevel phylogenetic placement," *Bioinformatics*, vol. 35, no. 7, pp. 1151–1158, 2019.

[55] S. A. Berger and A. Stamatakis, "Papara 2.0: a vectorized algorithm for probabilistic phylogeny-aware alignment extension," *Heidelberg Institute for Theoretical Studies*, 2012.

[56] I. Letunic and P. Bork, "Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation," *Nucleic acids research*, vol. 49, no. W1, pp. W293–W296, 2021.

[57] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic acids research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[58] T. Nakamura, K. D. Yamada, K. Tomii, and K. Katoh, "Parallelization of mafft for large-scale multiple sequence alignments," *Bioinformatics*, vol. 34, no. 14, pp. 2490–2492, 2018.

[59] D. M. Gohl, P. Vangay, J. Garbe, A. MacLean, A. Hauge, A. Becker, T. J. Gould, J. B. Clayton, T. J. Johnson, R. Hunter, *et al.*, "Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies," *Nature biotechnology*, vol. 34, no. 9, pp. 942–949, 2016.

[60] I. M. Bradley, A. J. Pinto, J. S. Guest, and G. Voordouw, "Design and evaluation of illumina miseq-compatible, 18s rrna gene-specific primers for improved characterization of mixed phototrophic communities," *Applied and Environmental Microbiology*, vol. 82, no. 19, pp. 5878–5891, 2016.

[61] M. G. Bakker, "A fungal mock community control for amplicon sequencing experiments," *Molecular ecology resources*, vol. 18, no. 3, pp. 541–556, 2018.

[62] I. Bista, G. R. Carvalho, M. Tang, K. Walsh, X. Zhou, M. Hajibabaei, S. Shokralla, M. Seymour, D. Bradley, S. Liu, *et al.*, "Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples," *Molecular Ecology Resources*, vol. 18, no. 5, pp. 1020–1034, 2018.

[63] C. Pavloudi, J. B. Kristoffersen, A. Oulas, M. De Troch, and C. Arvanitidis, "Sediment microbial taxonomic and functional diversity in a natural salinity gradient challenge remane's "species minimum" concept," *PeerJ*, vol. 5, p. e3687, 2017.

[64] I. Bista, G. Carvalho, K. Walsh, M. Seymour, M. Hajibabaei, D. Lallias, M. Christmas, and S. Creer, "Annual time-series analysis of aqueous edna reveals ecologically relevant dynamics of lake ecosystem biodiversity. nat. commun. 8, 14087," 2017.

[65] P. W. Harrison, B. Alako, C. Amid, A. Cerdeño-Tárraga, I. Cleland, S. Holt, A. Hussein, S. Jayathilaka, S. Kay, T. Keane, *et al.*, "The european nucleotide archive in 2018," *Nucleic acids research*, vol. 47, no. D1, pp. D84–D88, 2019.

[66] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[67] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "Blast+: architecture and applications," *BMC bioinformatics*, vol. 10, no. 1, pp. 1–9, 2009.

[68] S. Ratnasingham and P. D. Hebert, "Bold: The barcode of life data system (http://www. barcodinglife. org)," *Molecular ecology notes*, vol. 7, no. 3, pp. 355–364, 2007.

[69] F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn, "Swarm: robust and fast clustering method for amplicon-based studies," *PeerJ*, vol. 2, p. e593, 2014.

[70] S. I. Glassman and J. B. Martiny, "Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units," *MSphere*, vol. 3, no. 4, pp. e00148–18, 2018.

[71] P. Taberlet, E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev, "Towards next-generation biodiversity assessment using dna metabarcoding," *Molecular ecology*, vol. 21, no. 8, pp. 2045–2050, 2012.

[72] T. Schenekar, M. Schletterer, L. A. Lecaudey, and S. J. Weiss, "Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an edna fish assessment in the volga headwaters," *River Research and Applications*, vol. 36, no. 7, pp. 1004–1013, 2020.

[73] K. Cilleros, A. Valentini, L. Allard, T. Dejean, R. Etienne, G. Grenouillet, A. Iribar, P. Taberlet, R. Vigouroux, and S. Brosse, "Unlocking biodiversity and conservation studies in high-diversity environments using environmental dna (edna): A test with guianese freshwater fishes," *Molecular Ecology Resources*, vol. 19, no. 1, pp. 27–46, 2019.

[74] W. J. Kress, C. García-Robledo, M. Uriarte, and D. L. Erickson, "Dna barcodes for ecology, evolution, and conservation," *Trends in ecology & evolution*, vol. 30, no. 1, pp. 25–35, 2015.

[75] P. D. Hebert, S. Ratnasingham, and J. R. De Waard, "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. suppl_1, pp. S96–S99, 2003.

[76] V. Elbrecht and F. Leese, "Validation and development of coi metabarcoding primers for freshwater macroinvertebrate bioassessment," *Frontiers in Environmental Science*, vol. 5, p. 11, 2017.

[77] M. Miya, R. O. Gotoh, and T. Sado, "Mifish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental dna and other samples," *Fisheries Science*, pp. 1–32, 2020.

[78] S. Ekici, G. Pawlik, E. Lohmeyer, H.-G. Koch, and F. Daldal, "Biogenesis of cbb3-type cytochrome c oxidase in rhodobacter capsulatus," *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, vol. 1817, no. 6, pp. 898–910, 2012.

[79] S. Schimo, I. Wittig, K. M. Pos, and B. Ludwig, "Cytochrome c oxidase biogenesis and metallochaperone interactions: steps in the assembly pathway of a bacterial complex," *PLoS One*, vol. 12, no. 1, p. e0170037, 2017.

[80] H. Song, J. E. Buhay, M. F. Whiting, and K. A. Crandall, "Many species in one: Dna barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified," *Proceedings of the national academy of sciences*, vol. 105, no. 36, pp. 13486–13491, 2008.

[81] D. Bensasson, D.-X. Zhang, D. L. Hartl, and G. M. Hewitt, "Mitochondrial pseudogenes: evolution's misplaced witnesses," *Trends in ecology & evolution*, vol. 16, no. 6, pp. 314–321, 2001.

[82] M. Mioduchowska, M. J. Czyż, B. Gołdyn, J. Kur, and J. Sell, "Instances of erroneous dna barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"?," *PLoS One*, vol. 13, no. 6, p. e0199609, 2018.

[83] M. E. Siddall, F. M. Fontanella, S. C. Watson, S. Kvist, and C. Erséus, "Barcoding bamboozled by bacteria: convergence to metazoan mitochondrial primer targets by marine microbes," *Systematic Biology*, vol. 58, no. 4, pp. 445–451, 2009.

[84] C. Andújar, P. Arribas, D. W. Yu, A. P. Vogler, and B. C. Emerson, "Why the coi barcode should be the community dna metabarcode for the metazoa," 2018.

[85] P. Taberlet, A. Bonin, L. Zinger, and E. Coissac, "Analysis of bulk samples," in *Environmental DNA*, pp. 140–143, Oxford University Press.

[86] C. Yang, Y. Ji, X. Wang, C. Yang, and W. Y. Douglas, "Testing three pipelines for 18s rdna-based metabarcoding of soil faunal diversity," *Science China Life Sciences*, vol. 56, no. 1, pp. 73–81, 2013.

[87] C. Yang, X. Wang, J. A. Miller, M. de Blécourt, Y. Ji, C. Yang, R. D. Harrison, and W. Y. Douglas, "Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator," *Ecological Indicators*, vol. 46, pp. 379–389, 2014.

[88] R. A. Collins, J. Bakker, O. S. Wangensteen, A. Z. Soto, L. Corrigan, D. W. Sims, M. J. Genner, and S. Mariani, "Non-specific amplification compromises environmental dna metabarcoding with coi," *Methods in Ecology and Evolution*, vol. 10, no. 11, pp. 1985–2001, 2019.

[89] E. Aylagas, Á. Borja, X. Irigoien, and N. Rodríguez-Ezpeleta, "Benchmarking dna metabarcoding for biodiversity-based monitoring and assessment," *Frontiers in Marine Science*, vol. 3, p. 96, 2016.

[90] F. Sinniger, J. Pawlowski, S. Harii, A. J. Gooday, H. Yamamoto, P. Chevaldonné, T. Cedhagen, G. Carvalho, and S. Creer, "Worldwide analysis of sedimentary dna reveals major gaps in taxonomic knowledge of deep-sea benthos," *Frontiers in Marine Science*, vol. 3, p. 92, 2016.

[91] Q. Haenel, O. Holovachov, U. Jondelius, P. Sundberg, and S. J. Bourlat, "Ngs-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from hållö island, smögen, and soft mud from gullmarn fjord, sweden," *Biodiversity data journal*, no. 5, 2017.

[92] G. Bernard, J. S. Pathmanathan, R. Lannes, P. Lopez, and E. Bapteste, "Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery," *Genome biology and evolution*, vol. 10, no. 3, pp. 707–715, 2018.

[93] M. Jamy, R. Foster, P. Barbera, L. Czech, A. Kozlov, A. Stamatakis, G. Bending, S. Hilton, D. Bass, and F. Burki, "Long-read metabarcoding of the eukaryotic rdna operon to phylogenetically and taxonomically resolve environmental diversity," *Molecular ecology resources*, vol. 20, no. 2, pp. 429–443, 2020.

[94] L. Czech, P. Barbera, and A. Stamatakis, "Genesis and gappa: processing, analyzing and visualizing phylogenetic (placement) data," *Bioinformatics*, vol. 36, no. 10, pp. 3263–3265, 2020.

[95] J. L. Steenwyk, T. J. Buida III, Y. Li, X.-X. Shen, and A. Rokas, "Clipkit: A multiple sequence alignment trimming software for accurate phylogenomic inference," *PLoS biology*, vol. 18, no. 12, p. e3001007, 2020.

[96] D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, and L. S. Vinh, "Ufboot2: improving the ultrafast bootstrap approximation," *Molecular biology and evolution*, vol. 35, no. 2, pp. 518–522, 2018.

[97] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear, "Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era," *Molecular biology and evolution*, vol. 37, no. 5, pp. 1530–1534, 2020.

[98] H. Zafeiropoulos, A. Gioti, S. Ninidakis, A. Potirakis, S. Paragkamian, N. Angelova, A. Antoniou, T. Danis, E. Kaitetzidou, P. Kasapidis, *et al.*, "0s and 1s in marine molecular research: a regional hpc perspective," *GigaScience*, vol. 10, no. 8, p. giab053, 2021.

[99] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a web browser," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–10, 2011.

[100] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "Dada2: high-resolution sample inference from illumina amplicon data," *Nature methods*, vol. 13, no. 7, pp. 581–583, 2016.

[101] H. Zafeiropoulos, H. Q. Viet, K. Vasileiadou, A. Potirakis, C. Arvanitidis, P. Topalis, C. Pavloudi, and E. Pafilis, "Pema: a flexible pipeline for environmental dna metabarcoding analysis of the 16s/18s ribosomal rna, its, and coi marker genes," *GigaScience*, vol. 9, no. 3, p. giaa022, 2020.

[102] A. Antich, C. Palacin, O. S. Wangensteen, and X. Turon, "To denoise or to cluster, that is not the question: optimizing pipelines for coi metabarcoding and metaphylogeography," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–24, 2021.

[103] M. Obst, K. Exter, A. L. Allcock, C. Arvanitidis, A. Axberg, M. Bustamante, I. Cancio, D. Carreira-Flores, E. Chatzinikolaou, G. Chatzigeorgiou, *et al.*, "A marine biodiversity observation network for genetic monitoring of hard-bottom communities (arms-mbon)," *Frontiers in Marine Science*, vol. 7, p. 1031, 2020.

[104] S. Kamenova, "A flexible pipeline combining clustering and correction tools for prokaryotic and eukaryotic metabarcoding. peer community in ecology 1: 100043," 2020.

[105] S. Kumar, M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter, "Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated gc-coverage plots," *Frontiers in genetics*, vol. 4, p. 237, 2013.

[106] S. M. Dittami and E. Corre, "Detection of bacterial contaminants and hybrid sequences in the genome of the kelp saccharina japonica using taxoblast," *PeerJ*, vol. 5, p. e4073, 2017.

[107] G. De Simone, A. Pasquadibisceglie, R. Proietto, F. Polticelli, S. Aime, H. JM Op den Camp, and P. Ascenzi, "Contaminations in (meta) genome data: An open issue for the scientific community," *IUBMB life*, vol. 72, no. 4, pp. 698–705, 2020.

[108] M. Steinegger and S. L. Salzberg, "Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in genbank," *Genome biology*, vol. 21, no. 1, pp. 1–12, 2020.

[109] S. F. Gilbert, J. Sapp, and A. I. Tauber, "A symbiotic view of life: we have never been individuals," *The Quarterly review of biology*, vol. 87, no. 4, pp. 325–341, 2012.

[110] E. Salvucci, "Microbiome, holobiont and the net of life," *Critical reviews in microbiology*, vol. 42, no. 3, pp. 485–494, 2016.

[111] H. Weigand, A. J. Beermann, F. Čiampor, F. O. Costa, Z. Csabai, S. Duarte, M. F. Geiger, M. Grabowski, F. Rimet, B. Rulik, *et al.*, "Dna barcode reference libraries for the monitoring of aquatic biota in europe: Gap-analysis and recommendations for future work," *Science of the Total Environment*, vol. 678, pp. 499–524, 2019.

[112] G. Huys, T. Vanhoutte, M. Joossens, A. S. Mahious, E. De Brandt, S. Vermeire, and J. Swings, "Coamplification of eukaryotic dna with 16s rrna gene-based pcr primers: possible consequences for population fingerprinting of complex microbial communities," *Current microbiology*, vol. 56, no. 6, pp. 553–557, 2008.

[113] M. Delgado-Baquerizo, F. T. Maestre, P. B. Reich, T. C. Jeffries, J. J. Gaitan, D. Encinar, M. Berdugo, C. D. Campbell, and B. K. Singh, "Microbial diversity drives multifunctionality in terrestrial ecosystems," *Nature communications*, vol. 7, no. 1, pp. 1–8, 2016.

[114] A. Morris, K. Meyer, and B. Bohannan, "Linking microbial communities to ecosystem functions: what we can learn from genotype–phenotype mapping in organisms," *Philosophical Transactions of the Royal Society B*, vol. 375, no. 1798, p. 20190244, 2020.

[115] M. B. Biggs, G. L. Medlock, G. L. Kolling, and J. A. Papin, "Metabolic network modeling of microbial communities," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 7, no. 5, pp. 317–334, 2015.

[116] E. K. Hall, E. S. Bernhardt, R. L. Bier, M. A. Bradford, C. M. Boot, J. B. Cotner, P. A. Del Giorgio, S. E. Evans, E. B. Graham, S. E. Jones, *et al.*, "Understanding how

microbiomes influence the systems they inhabit," *Nature Microbiology*, vol. 3, no. 9, pp. 977–982, 2018.

[117] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nature reviews genetics*, vol. 7, no. 2, pp. 119–129, 2006.

[118] T. O. Delmont, C. Malandain, E. Prestat, C. Larose, J.-M. Monier, P. Simonet, and T. M. Vogel, "Metagenomic mining for microbiologists," *The ISME Journal*, vol. 5, no. 12, pp. 1837–1843, 2011.

[119] J. Raes and P. Bork, "Molecular eco-systems biology: towards an understanding of community function," *Nature Reviews Microbiology*, vol. 6, no. 9, pp. 693–699, 2008.

[120] R. H. Nilsson, S. Anslan, M. Bahram, C. Wurzbacher, P. Baldrian, and L. Tedersoo, "Mycobiome diversity: high-throughput sequencing and identification of fungi," *Nature Reviews Microbiology*, vol. 17, no. 2, pp. 95–109, 2019.

[121] S. Pesant, F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich, R. Troublé, *et al.*, "Open science resources for the discovery and analysis of tara oceans data," *Scientific data*, vol. 2, no. 1, pp. 1–16, 2015.

[122] J. A. Gilbert, J. K. Jansson, and R. Knight, "The earth microbiome project: successes and aspirations," *BMC biology*, vol. 12, no. 1, pp. 1–4, 2014.

[123] W.-S. Shu and L.-N. Huang, "Microbial diversity in extreme environments," *Nature Reviews Microbiology*, pp. 1–17, 2021.

[124] P. Yilmaz, R. Kottmann, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler, J. A. Gilbert, I. Karsch-Mizrachi, A. Johnston, G. Cochrane, *et al.*, "Minimum information about a marker gene sequence (mimarks) and minimum information about any (x) sequence (mixs) specifications," *Nature biotechnology*, vol. 29, no. 5, pp. 415–420, 2011.

[125] E. M. Wood-Charlson, D. Auberry, H. Blanco, M. I. Borkum, Y. E. Corilo, K. W. Davenport, S. Deshpande, R. Devarakonda, M. Drake, W. D. Duncan, *et al.*, "The national microbiome data collaborative: enabling microbiome science," *Nature Reviews Microbiology*, vol. 18, no. 6, pp. 313–314, 2020.

[126] P. Vangay, J. Burgin, A. Johnston, K. L. Beck, D. C. Berrios, K. Blumberg, S. Canon, P. Chain, J.-M. Chandonia, D. Christianson, *et al.*, "Microbiome metadata standards: Report of the national microbiome data collaborative's workshop and follow-on activities," *Msystems*, vol. 6, no. 1, pp. e01194–20, 2021.

[127] R. L. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, P. L. Buttigieg, N. Davies, D. Endresen, *et al.*, "Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies," *PloS one*, vol. 9, no. 3, p. e89606, 2014.

[128] P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, and C. J. Mungall, "The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation," *Journal of biomedical semantics*, vol. 7, no. 1, pp. 1–12, 2016.

[129] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[130] "The gene ontology resource: enriching a gold mine," *Nucleic Acids Research*, vol. 49, no. D1, pp. D325–D334, 2021.

[131] "IUPAC-IUBMB joint commission on biochemical nomenclature (JCBN) and nomenclature committee of IUBMB (NC-IUBMB), newsletter 1999," *Eur. J. Biochem.*, vol. 264, pp. 607–609, Sept. 1999.

[132] R. Caspi, R. Billington, I. M. Keseler, A. Kothari, M. Krummenacker, P. E. Midford, W. K. Ong, S. Paley, P. Subhraveti, and P. D. Karp, "The metacyc database of metabolic pathways and enzymes-a 2019 update," *Nucleic acids research*, vol. 48, no. D1, pp. D445–D453, 2020.

[133] C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, *et al.*, "Ncbi taxonomy: a comprehensive update on curation, resources and tools," *Database*, vol. 2020, 2020.

[134] A. C. Parte, J. S. Carbasse, J. P. Meier-Kolthoff, L. C. Reimer, and M. Göker, "List of prokaryotic names with standing in nomenclature (lpsn) moves to the dsmz," *International journal of systematic and evolutionary microbiology*, vol. 70, no. 11, p. 5607, 2020.

[135] A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, *et al.*, "Mgnify: the microbiome analysis resource in 2020," *Nucleic acids research*, vol. 48, no. D1, pp. D570–D578, 2020.

[136] I.-M. A. Chen, K. Chu, K. Palaniappan, A. Ratner, J. Huang, M. Huntemann, P. Hajek, S. Ritter, N. Varghese, R. Seshadri, *et al.*, "The img/m data management and analysis system v. 6.0: new tools and advanced capabilities," *Nucleic acids research*, vol. 49, no. D1, pp. D751–D763, 2021.

[137] A. Wilke, J. Bischof, T. Harrison, T. Brettin, M. D'Souza, W. Gerlach, H. Matthews, T. Paczian, J. Wilkening, E. M. Glass, *et al.*, "A restful api for accessing microbial community data for mg-rast," *PLoS computational biology*, vol. 11, no. 1, p. e1004008, 2015.

[138] R. J. Roberts, "Pubmed central: The genbank of the published literature," 2001.

[139] N. Harmston, W. Filsell, and M. P. Stumpf, "What the papers say: Text mining for genomics and systems biology," *Human genomics*, vol. 5, no. 1, pp. 1–13, 2010.

[140] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen, "The species and organisms resources for fast and accurate identification of taxonomic names in text," *PloS one*, vol. 8, no. 6, p. e65390, 2013.

[141] E. Pafilis, P. L. Buttigieg, B. Ferrell, E. Pereira, J. Schnetzer, C. Arvanitidis, and L. J. Jensen, "Extract: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation," *Database*, vol. 2016, 2016.

[142] C. Von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, "String: known and predicted protein–protein associations, integrated and transferred across organisms," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D433–D437, 2005.

[143] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. Von Mering, *et al.*, "String v9. 1: protein-protein interaction networks, with increased coverage and integration," *Nucleic acids research*, vol. 41, no. D1, pp. D808–D815, 2012.

[144] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér, "Data integration in the era of omics: current and future challenges," *BMC systems biology*, vol. 8, no. 2, pp. 1–10, 2014.

[145] K. D'Hondt, T. Kostic, R. McDowell, F. Eudes, B. K. Singh, S. Sarkar, M. Markakis, B. Schelkle, E. Maguin, and A. Sessitsch, "Microbiome innovations for a sustainable future," *Nature Microbiology*, vol. 6, no. 2, pp. 138–142, 2021.

[146] N. Conde-Pueyo, B. Vidiella, J. Sardanyés, M. Berdugo, F. T. Maestre, V. De Lorenzo, and R. Solé, "Synthetic biology for terraformation lessons from mars, earth, and the microbiome," *Life*, vol. 10, no. 2, p. 14, 2020.

[147] F. A. Baltoumas, S. Zafeiropoulou, E. Karatzas, M. Koutrouli, F. Thanati, K. Voutsadaki, M. Gkonta, J. Hotova, I. Kasionis, P. Hatzis, *et al.*, "Biomolecule and bioentity interaction databases in systems biology: A comprehensive review," *Biomolecules*, vol. 11, no. 8, p. 1245, 2021.

[148] L. C. Reimer, A. Vetcininova, J. S. Carbasse, C. Söhngen, D. Gleim, C. Ebeling, and J. Overmann, "Bac dive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis," *Nucleic acids research*, vol. 47, no. D1, pp. D631–D636, 2019.

[149] H. Shaaban, D. A. Westfall, R. Mohammad, D. Danko, D. Bezdan, E. Afshinnekoo, N. Segata, and C. E. Mason, "The microbe directory: An annotated, searchable inventory of microbes' characteristics," *Gates open research*, vol. 2, 2018.

[150] S. M. Kosina, A. M. Greiner, R. K. Lau, S. Jenkins, R. Baran, B. P. Bowen, and T. R. Northen, "Web of microbes (wom): a curated microbial exometabolomics database for linking chemistry and microbes," *BMC microbiology*, vol. 18, no. 1, pp. 1–10, 2018.

[151] Y. Tang, T. Dai, Z. Su, K. Hasegawa, J. Tian, L. Chen, and D. Wen, "A tripartite microbial-environment network indicates how crucial microbes influence the microbial community ecology," *Microbial ecology*, vol. 79, no. 2, pp. 342–356, 2020.

[152] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos, "A guide to conquer the biological network era using graph theory," *Frontiers in bioengineering and biotechnology*, vol. 8, p. 34, 2020.

[153] K. Li, J. Hu, T. Li, F. Liu, J. Tao, J. Liu, Z. Zhang, X. Luo, L. Li, Y. Deng, *et al.*, "Microbial abundance and diversity investigations along rivers: Current knowledge and future directions," *Wiley Interdisciplinary Reviews: Water*, vol. 8, no. 5, p. e1547, 2021.

[154] L. J. Jensen, "One tagger, many uses: Illustrating the power of ontologies in dictionary-based named entity recognition," *bioRxiv*, p. 067132, 2016.

[155] E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 49, no. D1, p. D10, 2021.

[156] E. Pafilis, S. P. Frankild, J. Schnetzer, L. Fanini, S. Faulwetter, C. Pavloudi, K. Vasileiadou, P. Leary, J. Hammock, K. Schulz, *et al.*, "Environments and eol: identification of environment ontology terms in text and the annotation of the encyclopedia of life," *Bioinformatics*, vol. 31, no. 11, pp. 1872–1874, 2015.

[157] S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, J. C. Sundaramurthi, J. Lee, M. Kandimalla, I.-M. A. Chen, N. C. Kyrpides, and T. Reddy, "Genomes online database (gold) v. 8: overview and updates," *Nucleic Acids Research*, vol. 49, no. D1, pp. D723–D733, 2021.

[158] J. de la Cuesta-Zuluaga, R. E. Ley, and N. D. Youngblut, "Struo: a pipeline for building custom databases for common metagenome profilers," *Bioinformatics*, vol. 36, no. 7, pp. 2314–2315, 2020.

[159] D. H. Parks, M. Chuvochina, P.-A. Chaumeil, C. Rinke, A. J. Mussig, and P. Hugenholtz, "A complete domain-to-species taxonomy for bacteria and archaea," *Nature biotechnology*, vol. 38, no. 9, pp. 1079–1086, 2020.

[160] L. Guillou, D. Bachar, S. Audic, D. Bass, C. Berney, L. Bittner, C. Boutte, G. Burgaud, C. De Vargas, J. Decelle, *et al.*, "The protist ribosomal reference database (pr2): a catalog of unicellular eukaryote small sub-unit rrna sequences with curated taxonomy," *Nucleic acids research*, vol. 41, no. D1, pp. D597–D604, 2012.

[161] J. Del Campo, M. Kolisko, V. Boscaro, L. F. Santoferrara, S. Nenarokov, R. Massana, L. Guillou, A. Simpson, C. Berney, C. de Vargas, *et al.*, "Eukref: phylogenetic curation of ribosomal rna to enhance understanding of eukaryotic diversity and distribution," *PLoS biology*, vol. 16, no. 9, p. e2005849, 2018.

[162] L. Suter, A. M. Polanowski, L. J. Clarke, J. A. Kitchener, and B. E. Deagle, "Capturing open ocean biodiversity: comparing environmental dna metabarcoding to the continuous plankton recorder," *Molecular ecology*, vol. 30, no. 13, pp. 3140–3157, 2021.

[163] M. Leray, S.-L. Ho, I.-J. Lin, and R. J. Machida, "Midori server: a webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database," *Bioinformatics*, vol. 34, no. 21, pp. 3753–3754, 2018.

[164] C. Pavloudi, A. Oulas, K. Vasileiadou, G. Kotoulas, M. De Troch, M. W. Friedrich, and C. Arvanitidis, "Diversity and abundance of sulfate-reducing microorganisms in a mediterranean lagoonal complex (amvrakikos gulf, ionian sea) derived from dsrb gene," *Aquatic Microbial Ecology*, vol. 79, no. 3, pp. 209–219, 2017.

[165] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak, "A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts," *PLoS computational biology*, vol. 14, no. 2, p. e1005962, 2018.

[166] C. Ferguson, D. Araújo, L. Faulk, Y. Gou, A. Hamelers, Z. Huang, M. Ide-Smith, M. Levchenko, N. Marinos, R. Nambiar, *et al.*, "Europe pmc in 2020," *Nucleic acids research*, vol. 49, no. D1, pp. D1507–D1514, 2021.

[167] E. Karatzas, F. A. Baltoumas, N. A. Panayiotou, R. Schneider, and G. A. Pavlopoulos, "Arena3dweb: Interactive 3d visualization of multilayered networks," *Nucleic Acids Research*, 2021.

[168] F. A. Baltoumas, S. Zafeiropoulou, E. Karatzas, S. Paragkamian, F. Thanati, I. Iliopoulos, A. G. Eliopoulos, R. Schneider, L. J. Jensen, E. Pafilis, *et al.*, "Onthefly2. 0: a text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis," *bioRxiv*, 2021.

[169] F. Thanati, E. Karatzas, F. Baltoumas, D. J. Stravopodis, A. G. Eliopoulos, and G. Pavlopoulos, "Flame: a web tool for functional and literature enrichment analysis of multiple gene lists," *bioRxiv*, 2021.

[170] J. Zoppi, J.-F. Guillaume, M. Neunlist, and S. Chaffron, "Mibiomics: an interactive web application for multi-omics data exploration and integration," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–14, 2021.

[171] L. Sinclair, U. Z. Ijaz, L. J. Jensen, M. J. Coolen, C. Gubry-Rangin, A. Chroňáková, A. Oulas, C. Pavloudi, J. Schnetzer, A. Weimann, *et al.*, "Seqenv: linking sequences to environments through text mining," *PeerJ*, vol. 4, p. e2690, 2016.

[172] C.-X. Xue, H. Lin, X.-Y. Zhu, J. Liu, Y. Zhang, G. Rowley, J. D. Todd, M. Li, and X.-H. Zhang, "Diting: a pipeline to infer and compare biogeochemical pathways from metagenomic and metatranscriptomic data," *Frontiers in microbiology*, p. 2118, 2021.

[173] E. Klipp, W. Liebermeister, C. Wierling, and A. Kowald, *Systems biology: a textbook*. John Wiley & Sons, 2016.

[174] P. Kohl, E. J. Crampin, T. Quinn, and D. Noble, "Systems biology: an approach," *Clinical Pharmacology & Therapeutics*, vol. 88, no. 1, pp. 25–33, 2010.

[175] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: systems biology," *Annual review of genomics and human genetics*, vol. 2, no. 1, pp. 343–372, 2001.

[176] R. A. Quinn, J. A. Navas-Molina, E. R. Hyde, S. J. Song, Y. Vázquez-Baeza, G. Humphrey, J. Gaffney, J. J. Minich, A. V. Melnik, J. Herschend, J. DeReus, A. Durant, R. J. Dutton, M. Khosroheidari, C. Green, R. da Silva, P. C. Dorrestein, and R. Knight, "From sample to multi-omics conclusions in under 48 hours. msystems 1: e00038-16," *Crossref, Medline*, 2016.

[177] D. Noble, *The music of life: biology beyond genes*. Oxford University Press, 2008.

[178] B. Ø.. Palsson, *Systems biology*. Cambridge university press, 2015.

[179] J. R. Schramski, A. I. Dell, J. M. Grady, R. M. Sibly, and J. H. Brown, "Metabolic theory predicts whole-ecosystem properties," *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, pp. 2617–2622, 2015.

[180] I. Thiele and B. Ø.. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction," *Nature protocols*, vol. 5, no. 1, p. 93, 2010.

[181] D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil, "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities," *Nucleic acids research*, vol. 46, no. 15, pp. 7542–7553, 2018.

[182] B. Ø.. Palsson, "Metabolic systems biology," *FEBS letters*, vol. 583, no. 24, pp. 3900–3904, 2009.

[183] S. S. Shishvan, A. Vigliotti, and V. S. Deshpande, "The homeostatic ensemble for cells," *Biomechanics and Modeling in Mechanobiology*, vol. 17, no. 6, pp. 1631–1662, 2018.

[184] A. Cakmak, X. Qi, A. E. Cicek, I. Bederman, L. Henderson, M. Drumm, and G. Ozsoyoglu, "A new metabolomics analysis technique: steady-state metabolic network dynamics analysis," *Journal of bioinformatics and computational biology*, vol. 10, no. 01, p. 1240003, 2012.

[185] M. Lularevic, A. J. Racher, C. Jaques, and A. Kiparissides, "Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions," *Biotechnology and bioengineering*, vol. 116, no. 9, pp. 2339–2352, 2019.

[186] J. D. Orth, I. Thiele, and B. Ø.. Palsson, "What is flux balance analysis?," *Nature biotechnology*, vol. 28, no. 3, pp. 245–248, 2010.

[187] J. Schellenberger and B. Ø.. Palsson, "Use of randomized sampling for analysis of metabolic networks," *Journal of biological chemistry*, vol. 284, no. 9, pp. 5457–5461, 2009.

[188] R. L. Smith, "Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions," *Operations Research*, vol. 32, no. 6, pp. 1296–1308, 1984.

[189] D. E. Kaufman and R. L. Smith, "Direction choice for accelerated convergence in hit-and-run sampling," *Operations Research*, vol. 46, no. 1, pp. 84–95, 1998.

[190] P. A. Saa and L. K. Nielsen, "ll-ACHRB: a scalable algorithm for sampling the feasible solution space of metabolic networks," *Bioinform.*, vol. 32, no. 15, pp. 2330–2337, 2016.

[191] H. S. Haraldsdóttir, B. Cousins, I. Thiele, R. M. Fleming, and S. Vempala, "CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models," *Bioinformatics*, vol. 33, no. 11, pp. 1741–1743, 2017.

[192] L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdóttir, J. Wachowiak, S. M. Keating, V. Vlasov, *et al.*, "Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0," *Nature protocols*, vol. 14, no. 3, pp. 639–702, 2019.

[193] M. MacGillivray, A. Ko, E. Gruber, M. Sawyer, E. Almaas, and A. Holder, "Robust analysis of fluxes in genome-scale metabolic pathways," *Scientific Reports*, vol. 7, 12 2017.

[194] S. Fallahi, H. J. Skaug, and G. Alendal, "A comparison of Monte Carlo sampling methods for metabolic network models," *PLOS One*, vol. 15, no. 7, p. e0235393, 2020.

[195] D. B. Bernstein, F. E. Dewhirst, and D. Segre, "Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome," *Elife*, vol. 8, p. e39733, 2019.

[196] O. Perez-Garcia, G. Lear, and N. Singhal, "Metabolic network modeling of microbial interactions in natural and engineered environmental systems," *Frontiers in microbiology*, vol. 7, p. 673, 2016.

[197] L. Lovász, R. Kannan, and M. Simonovits, "Random walks and an $O^*(n^5)$ volume algorithm for convex bodies," *Random Structures and Algorithms*, vol. 11, pp. 1–50, 1997.

[198] L. Lovász and S. Vempala, "Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithms," *J. Computer & System Sciences*, vol. 72, pp. 392–417, 2006.

[199] B. Cousins and S. Vempala, "A practical volume algorithm," *Mathematical Programming Computation*, vol. 8, no. 2, pp. 133–160, 2016.

[200] R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann, "Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles," *Journal of the American Mathematical Society*, vol. 23, no. 2, pp. 535–561, 2010.

[201] F. John, "Extremum Problems with Inequalities as Subsidiary Conditions," in *Traces and Emergence of Nonlinear Programming* (G. Giorgi and T. H. Kjeldsen, eds.), pp. 197–215, Basel: Springer, 2014.

[202] D. Bertsimas and S. Vempala, "Solving convex programs by random walks," *J. ACM*, vol. 51, p. 540–556, July 2004.

[203] A. T. Kalai and S. Vempala, "Simulated annealing for convex optimization," *Mathematics of Operations Research*, vol. 31, no. 2, pp. 253–266, 2006.

[204] A. Laddha and S. Vempala, "Convergence of Gibbs Sampling: Coordinate Hit-and-Run Mixes Fast," 2020.

[205] H. Narayanan and P. Srivastava, "On the mixing time of coordinate hit-and-run," 2020.

[206] A. Chalkis, I. Z. Emiris, and V. Fisikopoulos, "Practical volume estimation by a new annealing schedule for cooling convex bodies," 2019.

[207] Y. T. Lee and S. S. Vempala, "Convergence rate of riemannian hamiltonian monte carlo and faster polytope volume computation," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, (New York, NY, USA), p. 1115–1121, Association for Computing Machinery, 2018.

[208] Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu, "Fast mcmc sampling algorithms on polytopes," *Journal of Machine Learning Research*, vol. 19, no. 55, pp. 1–86, 2018.

[209] E. Gryazina and B. Polyak, "Random sampling: Billiard walk algorithm," *European Journal of Operational Research*, vol. 238, no. 2, pp. 497 – 504, 2014.

[210] A. B. Dieker and S. S. Vempala, "Stochastic billiards for sampling from the boundary of a convex set," *Mathematics of Operations Research*, vol. 40, no. 4, pp. 888–901, 2015.

[211] A. Chalkis, I. Z. Emiris, and V. Fisikopoulos, "Practical volume estimation of zonotopes by a new annealing schedule for cooling convex bodies," in *Mathematical Software – ICMS 2020* (A. M. Bigatti, J. Carette, J. H. Davenport, M. Joswig, and T. de Wolff, eds.), (Cham), pp. 212–221, Springer International Publishing, 2020.

[212] A. Chevallier, S. Pion, and F. Cazals, "Hamiltonian Monte Carlo with boundary reflections, and application to polytope volume calculations," Research Report RR-9222, INRIA Sophia Antipolis, France, 2018.

[213] V. Roy, "Convergence Diagnostics for Markov Chain Monte Carlo," *Annual Review of Statistics and Its Application*, vol. 7, no. 1, pp. 387–412, 2020.

[214] C. J. Geyer, "Practical Markov chain Monte Carlo," *Statist. Sci.*, vol. 7, pp. 473–483, 11 1992.

[215] A. Gelman and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–472, 1992. Publisher: Institute of Mathematical Statistics.

[216] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.

[217] B. Cousins, *Efficient high-dimensional sampling and integration.* PhD thesis, Georgia Institute of Technology, Georgia, U.S.A., 2017.

[218] L. Calès, A. Chalkis, I. Z. Emiris, and V. Fisikopoulos, "Practical Volume Computation of Structured Convex Bodies, and an Application to Modeling Portfolio Dependencies and Financial Crises," in *34th International Symposium on Computational Geometry (SoCG 2018)* (B. Speckmann and C. D. Tóth, eds.), vol. 99 of *LIPIcs*, (Dagstuhl, Germany), pp. 19:1–19:15, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.

[219] S. Bordel, R. Agren, and J. Nielsen, "Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes," *PLOS Computational Biology*, vol. 6, pp. 1–13, 07 2010.

[220] S. Artstein-Avidan, H. Kaplan, and M. Sharir, "On radial isotropic position: Theory and algorithms," 2020.

[221] G. H.Golub and C. F. V. Loan, *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, 2013.

[222] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. Ø.. Palsson, and N. E. Lewis, "Bigg models: A platform for integrating, standardizing and sharing genome-scale models," *Nucleic acids research*, vol. 44, no. D1, pp. D515–D522, 2016.

[223] J. F. Jadebeck, A. Theorell, S. Leweke, and K. Noh, "Hops: high-performance library for non uniform sampling of convex constrained models," *Bioinformatics*, 2020.

[224] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 401–407, 1996.

[225] G. Guennebaud, B. Jacob, *et al.*, *Eigen v3*, 2010.

[226] A. Chalkis and V. Fisikopoulos, "volesti: Volume approximation and sampling for convex polytopes in R," 2020. https://github.com/GeomScale/volume_approximation.

[227] A. Noronha, J. Modamio, Y. Jarosz, E. Guerard, N. Sompairac, G. Preciat, A. D. Daníelsdóttir, M. Krecke, D. Merten, H. S. Haraldsdóttir, A. Heinken, L. Heirendt, S. Magnúsdóttir, D. A. Ravcheev, S. Sahoo, P. Gawron, L. Friscioni, B. Garcia, M. Prendergast, A. Puente, M. Rodrigues, A. Roy, M. Rouquaya, L. Wiltgen, A. Žagare, E. John, M. Krueger, I. Kuperstein, A. Zinovyev, R. Schneider, R. M. T. Fleming, and I. Thiele, "The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease," *Nucleic Acids Research*, vol. 47, pp. D614–D624, 10 2018.

[228] M. Heinonen, M. Osmala, H. Mannerström, J. Wallenius, S. Kaski, J. Rousu, and H. Lähdesmäki, "Bayesian metabolic flux analysis reveals intracellular flux couplings," *Bioinformatics*, vol. 35, no. 14, pp. i548–i557, 2019.

[229] NOAA, "How much water is in the ocean?," 2021. [Online; accessed 07-January-2022].

[230] J. A. Estes, M. Heithaus, D. J. McCauley, D. B. Rasher, and B. Worm, "Megafaunal impacts on structure and function of ocean ecosystems," *Annual Review of Environment and Resources*, vol. 41, pp. 83–116, 2016.

[231] K. R. Arrigo, "Marine microorganisms and global nutrient cycles," *Nature*, vol. 437, no. 7057, pp. 349–355, 2005.

[232] F. Boero and E. Bonsdorff, "A conceptual framework for marine biodiversity and ecosystem functioning," *Marine Ecology*, vol. 28, pp. 134–145, 2007.

[233] L. M. Beal, W. P. De Ruijter, A. Biastoch, and R. Zahn, "On the role of the agulhas system in ocean circulation and climate," *Nature*, vol. 472, no. 7344, pp. 429–436, 2011.

[234] K. Remoundou, P. Koundouri, A. Kontogianni, P. A. Nunes, and M. Skourtos, "Valuation of natural marine ecosystems: an economic perspective," *environmental science & policy*, vol. 12, no. 7, pp. 1040–1051, 2009.

[235] H.-O. Pörtner, D. C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, and N. Weyer, "The ocean and cryosphere in a changing climate," 2019.

[236] E. Sala and N. Knowlton, "Global marine biodiversity trends," *Annu. Rev. Environ. Resour.*, vol. 31, pp. 93–122, 2006.

[237] T. Tonon and D. Eveillard, "Marine systems biology," *Frontiers in genetics*, vol. 6, p. 181, 2015.

[238] H. M. Dionisi, M. Lozada, and N. L. Olivera, "Bioprospection of marine microorganisms: biotechnological applications and methods," *Revista argentina de microbiología*, vol. 44, no. 1, pp. 49–60, 2012.

[239] J. H. Tidwell and G. L. Allan, "Fish as food: aquaculture's contribution," *EMBO reports*, vol. 2, no. 11, pp. 958–963, 2001.

[240] G. Carvalho and L. Hauser, "Molecular genetics and the stock concept in fisheries," in *Molecular genetics in fisheries*, pp. 55–79, Springer, 1995.

[241] A. K. Sakai, F. W. Allendorf, J. S. Holt, D. M. Lodge, J. Molofsky, K. A. With, S. Baughman, R. J. Cabin, J. E. Cohen, N. C. Ellstrand, *et al.*, "The population biology of invasive species," *Annual review of ecology and systematics*, vol. 32, no. 1, pp. 305–332, 2001.

[242] G. A. Begg and J. R. Waldman, "An holistic approach to fish stock identification," *Fisheries research*, vol. 43, no. 1-3, pp. 35–44, 1999.

[243] M. Loreau, "Biodiversity and ecosystem functioning: recent theoretical advances," *Oikos*, vol. 91, no. 1, pp. 3–17, 2000.

[244] M. C. Leal, J. Puga, J. Serodio, N. C. Gomes, and R. Calado, "Trends in the discovery of new marine natural products from invertebrates over the last two decades–where and what are we bioprospecting?," *PLoS One*, vol. 7, no. 1, p. e30580, 2012.

[245] J. Norberg, D. P. Swaney, J. Dushoff, J. Lin, R. Casagrandi, and S. A. Levin, "Phenotypic diversity and ecosystem functioning in changing environments: a theoretical framework," *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11376–11381, 2001.

[246] E. R. Mardis, "Next-generation dna sequencing methods," *Annu. Rev. Genomics Hum. Genet.*, vol. 9, pp. 387–402, 2008.

[247] J. Kulski, *Next generation sequencing: advances, applications and challenges*. BoD–Books on Demand, 2016.

[248] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, 2016.

[249] J. G. Bundy, M. P. Davey, and M. R. Viant, "Environmental metabolomics: a critical review and future perspectives," *Metabolomics*, vol. 5, no. 1, pp. 3–21, 2009.

[250] V. Cahais, P. Gayral, G. Tsagkogeorga, J. Melo-Ferreira, M. Ballenghien, L. Weinert, Y. Chiari, K. Belkhir, V. Ranwez, and N. Galtier, "Reference-free transcriptome assembly in non-model animals from next-generation sequencing data," *Molecular ecology resources*, vol. 12, no. 5, pp. 834–845, 2012.

[251] N. A. Baird, P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson, "Rapid snp discovery and genetic mapping using sequenced rad markers," *PloS one*, vol. 3, no. 10, p. e3376, 2008.

[252] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in rna-seq: a matter of depth," *Genome research*, vol. 21, no. 12, pp. 2213–2223, 2011.

[253] J. E. Goldford, N. Lu, D. Bajić, S. Estrela, M. Tikhonov, A. Sanchez-Gorostiaga, D. Segrè, P. Mehta, and A. Sanchez, "Emergent simplicity in microbial community assembly," *Science*, vol. 361, no. 6401, pp. 469–474, 2018.

[254] I. Merelli, H. Pérez-Sánchez, S. Gesing, and D. D'Agostino, "Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives," *BioMed research international*, vol. 2014, 2014.

[255] J.-i. Sohn and J.-W. Nam, "The present and future of de novo whole-genome assembly," *Briefings in bioinformatics*, vol. 19, no. 1, pp. 23–40, 2018.

[256] C. S. Greene, J. Tan, M. Ung, J. H. Moore, and C. Cheng, "Big data bioinformatics," *Journal of cellular physiology*, vol. 229, no. 12, pp. 1896–1900, 2014.

[257] S. Pal, S. Mondal, G. Das, S. Khatua, and Z. Ghosh, "Big data in biology: The hope and present-day challenges in it," *Gene Reports*, p. 100869, 2020.

[258] S. Lampa, M. Dahlö, P. I. Olason, J. Hagberg, and O. Spjuth, "Lessons learned from implementing a national infrastructure in sweden for storage and analysis of next-generation sequencing data," *Gigascience*, vol. 2, no. 1, pp. 2047–217X, 2013.

[259] T. Sterling, M. Brodowicz, and M. Anderson, *High performance computing: modern systems and practices*. Morgan Kaufmann, 2017.

[260] Wikipedia contributors, "Supercomputing in europe — Wikipedia, the free encyclopedia," 2021. [Online; accessed 7-January-2022].

[261] E. Lindahl, "The scientific case for computing in europe 2018-2026," 2018.

[262] L. Candela, D. Castelli, and P. Pagano, "Virtual research environments: an overview and a research agenda. data sci. j," *J*, vol. 12, pp. 75–81, 2013.

[263] H. Zafeiropoulos, A. Gioti, S. Ninidakis, A. Potirakis, S. Paragkamian, N. Angelova, A. Antoniou, T. Danis, E. Kaitetzidou, P. Kasapidis, J. B. Kristoffersen, V. Papadogiannis, C. Pavloudi, Q. V. Ha, J. Lagnel, N. Pattakos, G. Perantinos, D. Sidirokastritis, P. Vavilis, G. Kotoulas, T. Manousaki, E. Sarropoulou, C. S. Tsigenopoulos, C. Arvanitidis, A. Magoulas, and E. Pafilis, "The IMBBC HPC facility: history, configuration, usage statistics and related activities." HZ and NG contributed equally at this work. Correspondence to: E. Pafilis; pafilis@hcmr.gr, Apr. 2021.

[264] J. J. Dongarra, P. Luszczek, and A. Petitet, "The linpack benchmark: past, present and future," *Concurrency and Computation: practice and experience*, vol. 15, no. 9, pp. 803–820, 2003.

[265] T. Castrignanò, S. Gioiosa, T. Flati, M. Cestari, E. Picardi, M. Chiara, M. Fratelli, S. Amente, M. Cirilli, M. A. Tangaro, *et al.*, "Elixir-it hpc@ cineca: high performance computing resources for the bioinformatics community," *BMC bioinformatics*, vol. 21, no. 10, pp. 1–17, 2020.

[266] J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko, "Stacks: an analysis tool set for population genomics," *Molecular ecology*, vol. 22, no. 11, pp. 3124–3140, 2013.

[267] C. Varsos, T. Patkos, A. Oulas, C. Pavloudi, A. Gougousis, U. Z. Ijaz, I. Filiopoulou, N. Pattakos, E. V. Berghe, A. Fernández-Guerra, *et al.*, "Optimized r functions for analysis of ecological community data using the r virtual laboratory (rvlab)," *Biodiversity data journal*, no. 4, 2016.

[268] K. E. Klymus, N. T. Marshall, and C. A. Stepien, "Environmental dna (edna) metabarcoding assays to detect invasive invertebrate species in the great lakes," *PloS one*, vol. 12, no. 5, p. e0177643, 2017.

[269] M. Bariche, S. A. Al-Mabruk, M. A. Ateş, A. Büyük, F. Crocetta, M. Dritsas, D. Edde, A. Fortič, E. Gavriil, V. Gerovasileiou, *et al.*, "New alien mediterranean biodiversity records (march 2020)," *Mediterranean Marine Science*, vol. 21, no. 1, pp. 129–145, 2020.

[270] S. Katsanevakis, M. Coll, C. Piroddi, J. Steenbeek, F. Ben Rais Lasram, A. Zenetos, and A. C. Cardoso, "Invading the mediterranean sea: biodiversity patterns shaped by human activities," *Frontiers in Marine Science*, vol. 1, p. 32, 2014.

[271] M. Pauletto, T. Manousaki, S. Ferraresso, M. Babbucci, A. Tsakogiannis, B. Louro, N. Vitulo, V. H. Quoc, R. Carraro, D. Bertotto, *et al.*, "Genomic analysis of sparus aurata reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish," *Communications biology*, vol. 1, no. 1, pp. 1–13, 2018.

[272] E. Sarropoulou, A. Y. Sundaram, E. Kaitetzidou, G. Kotoulas, G. D. Gilfillan, N. Papandroulakis, C. C. Mylonas, and A. Magoulas, "Full genome survey and dynamics of gene expression in the greater amberjack seriola dumerili," *GigaScience*, vol. 6, no. 12, p. gix108, 2017.

[273] A. Tsakogiannis, T. Manousaki, V. Anagnostopoulou, M. Stavroulaki, and E. T. Apostolaki, "The importance of genomics for deciphering the invasion success of the seagrass halophila stipulacea in the changing mediterranean sea," *Diversity*, vol. 12, no. 7, p. 263, 2020.

[274] T. Danis, A. Tsakogiannis, J. B. Kristoffersen, D. Golani, D. Tsaparis, P. Kasapidis, G. Kotoulas, A. Magoulas, C. S. Tsigenopoulos, and T. Manousaki, "Building a high-quality reference genome assembly for the the eastern mediterranean sea invasive sprinter lagocephalus sceleratus (tetraodontiformes, tetraodontidae)," *Biorxiv*, 2020.

[275] N. Angelova, T. Danis, L. Jacques, C. Tsigenopoulos, and T. Manousaki, "SnakeCube: containerized and automated next- generation sequencing (NGS) pipelines for genome analyses in HPC environments," Apr. 2021.

[276] P. Natsidis, A. Tsakogiannis, P. Pavlidis, C. S. Tsigenopoulos, and T. Manousaki, "Phylogenomics investigation of sparids (teleostei: Spariformes) using high-quality proteomes highlights the importance of taxon sampling," *Communications biology*, vol. 2, no. 1, pp. 1–10, 2019.

[277] M. Papadaki, E. Kaitetzidou, C. C. Mylonas, and E. Sarropoulou, "Non-coding rna expression patterns of two different teleost gonad maturation stages," *Marine Biotechnology*, vol. 22, no. 5, pp. 683–695, 2020.

[278] R. Warwick and P. Somerfield, "All animals are equal, but some animals are more equal than others," *Journal of Experimental Marine Biology and Ecology*, vol. 366, no. 1-2, pp. 184–186, 2008.

[279] C. D. Arvanitidis, R. M. Warwick, P. J. Somerfield, C. Pavloudi, E. Pafilis, A. Oulas, G. Chatzigeorgiou, V. Gerovasileiou, T. Patkos, N. Bailly, *et al.*, "Research infrastructures offer capacity to address scientific questions never attempted before: Are all taxa equal?," tech. rep., PeerJ Preprints, 2018.

[280] L. Vandepitte, B. Vanhoorne, W. Decock, S. Vranken, T. Lanssens, S. Dekeyzer, K. Verfaille, T. Horton, A. Kroh, F. Hernandez, *et al.*, "A decade of the world register of marine species–general insights and experiences from the data management team: Where are we, what have we learned and how can we continue?," *PLoS One*, vol. 13, no. 4, p. e0194599, 2018.

[281] A. Gioti, R. Siaperas, E. Nikolaivits, G. Le Goff, J. Ouazzani, G. Kotoulas, and E. Topakas, "Draft genome sequence of a cladosporium species isolated from the mesophotic ascidian didemnum maculosum," *Microbiology resource announcements*, vol. 9, no. 18, pp. e00311–20, 2020.

[282] E. Nikolaivits, R. Siaperas, A. Agrafiotis, J. Ouazzani, A. Magoulas, A. Gioti, and E. Topakas, "Functional and transcriptomic investigation of laccase activity in the presence of pcb29 identifies two novel enzymes and the multicopper oxidase repertoire of a marine-derived fungus," *Science of The Total Environment*, vol. 775, p. 145818, 2021.

[283] L. Dagum and R. Menon, "Openmp: an industry standard api for shared-memory programming," *IEEE computational science and engineering*, vol. 5, no. 1, pp. 46–55, 1998.

[284] P. D. Vouzis and N. V. Sahinidis, "Gpu-blast: using graphics processors to accelerate protein sequence alignment," *Bioinformatics*, vol. 27, no. 2, pp. 182–188, 2011.

[285] M. S. Nobile, P. Cazzaniga, A. Tangherloni, and D. Besozzi, "Graphics processing units in bioinformatics, computational biology and systems biology," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 870–885, 2017.

[286] P. Mell, T. Grance, *et al.*, "The nist definition of cloud computing," 2011.

[287] B. Langmead and A. Nellore, "Cloud computing for genomic data analysis and collaboration," *Nature Reviews Genetics*, vol. 19, no. 4, pp. 208–219, 2018.

[288] M. Dahlö, D. G. Scofield, W. Schaal, and O. Spjuth, "Tracking the ngs revolution: managing life science research on shared high-performance computing clusters," *GigaScience*, vol. 7, no. 5, p. giy028, 2018.

[289] J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, and L. J. Jensen, "Compartments: unification and visualization of protein subcellular localization evidence," *Database*, vol. 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** (2020 - ongoing)
  Role: technical support & principal investigator
  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.

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*

Publications

- **PREGO:**
  Zafeiropoulos, H.,

- **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

- **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

- **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

- **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

- **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