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

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

Here are some thoughts of mine for the rest of the committe!
Prof Faust
Prof. Elias Tsigaridas
Prof. Klappa
Prof L.J.J

Preface

Haris Zafeiropoulos

Contents

Abstract

Περίληψη

Και στα ελληνικά

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

NGS	Next Generation Sequencing
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

Chapter 1

Introduction

1.1 Microbial communities: structure & function

Microbes, i.e. Bacteria, Archaea and small Eukaryotes such as protoza, are omnipresent and impact global ecosystem functions [5] through their abundance [6], versatility [7] and interactions [8].

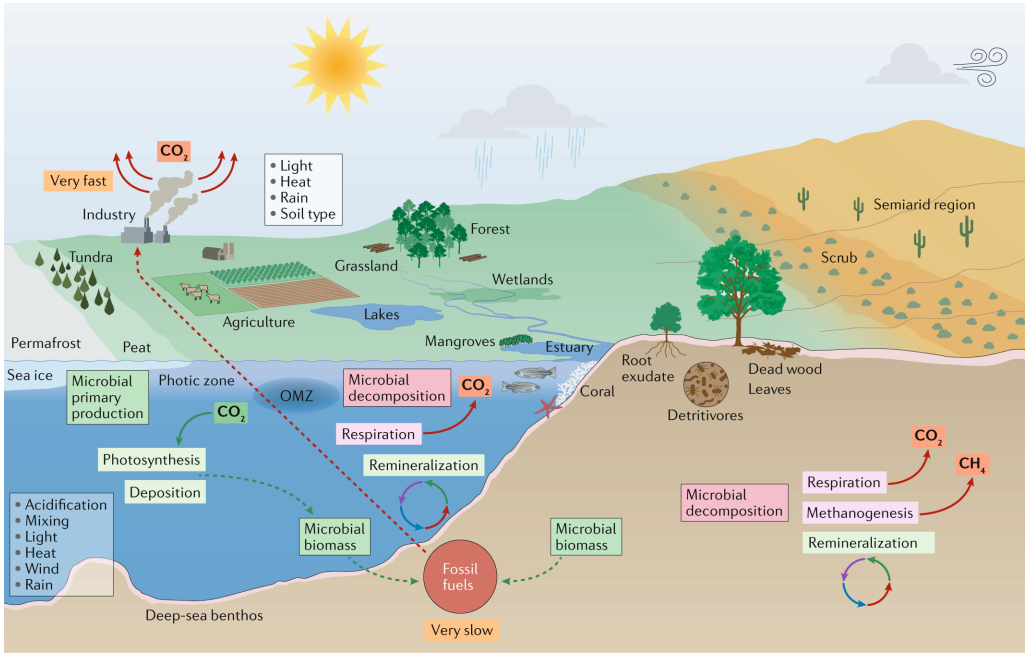

FIGURE 1.1: Marine icrobial communities contribute to CO_2 sequestration, nutrients recycle and thus to the release of CO_2 to the atmosphere. Soil microbial communities decomposers organic matter and release nutrients in the soil from [1] doi: 10.1038/s41579-019-0222-5, under Creative Commons Attribution 4.0 International License

1.1.1 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]

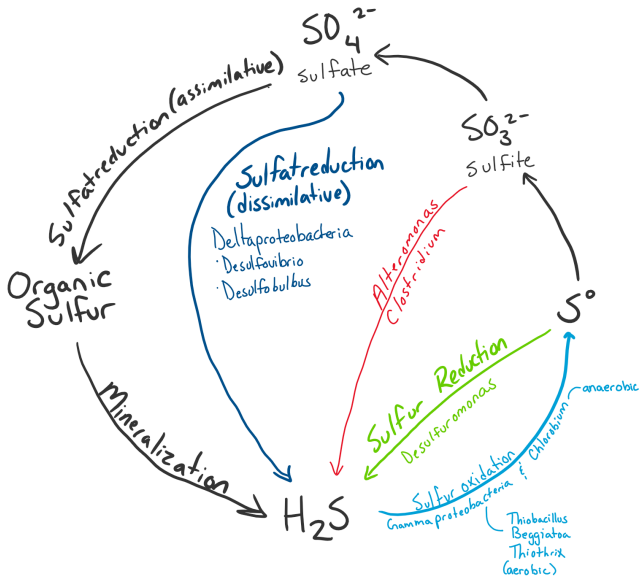

FIGURE 1.2: Sulfur cycle. Figure taken from [2]

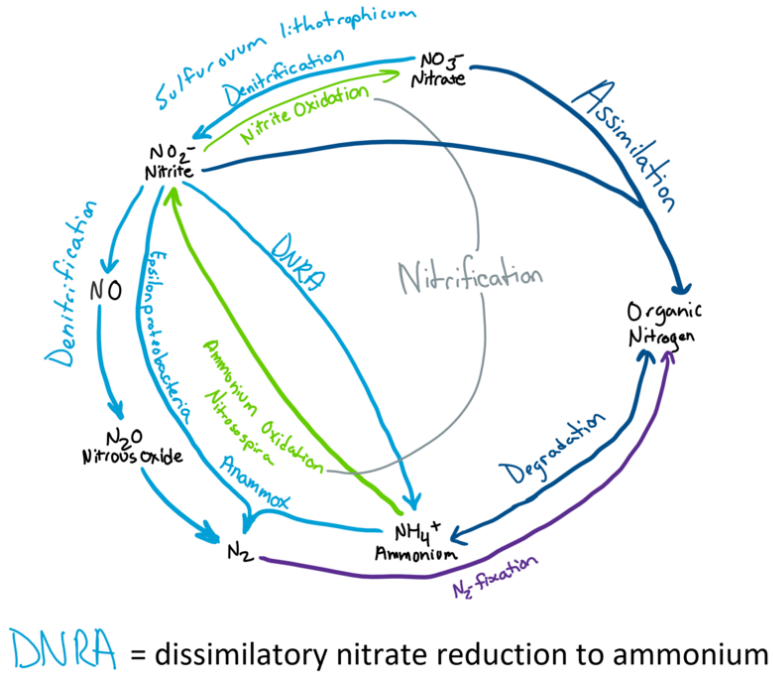

FIGURE 1.3: Nitrogen cycle. Figure taken from [3]

1.1.2 Microbial interactions: unravelling the microbiome

1.2 High Throughput Sequencing approaches and bioinformatics challenges

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" [9] provided the foundation for FAIR data management in the domain. These best practices advocated using community standards for contextual provenance and metadata at all stages of the research data life cycle.

Alongside archived sequence data, access to comprehensive metadata is important to contextualise where the data originated. On submission, submitters are given the option to provide details regarding when, where and how their samples were collected with the opportunity to align provided metadata against community developed standards where possible. However, challenges associated with metadata deposition mean submitters do not always provide comprehensive metadata - these challenges can range from: lack of training and outreach resulting in submitters not fully understanding the importance of metadata and how to comply with standards; as well as the trade-offs for the archives

to provide complex and thorough validation vs simple user interfaces to ensure both compliance and submission are as easy as possible. For the ENA, extensive documentation exists on how to submit data which both encourages compliance with metadata standards and provides separate submission guidelines for different data types - usage of the documentation can mitigate common errors and often aid first-time submitters but does not reach the full user-base.

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

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

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

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

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

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

- Metadata of the assembled data
- Metadata to accompany the taxonomic inventories

1.3.2 Ontologies & databases: the corner stone of mordern biology

Databases

- GenBank, ENA
- repositories such as MGnify
- PubMed

Ontologies:

- ENVO
- NCBI Taxonomy
- Gene Ontology
- Uniprot
- KEGG

1.4 Metabolic modeling at the omics era

1.4.1 Genome-scale metabolic model analysis

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 computational.

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

2.1.1 Introduction

Publication relative to this chapter: [10].

2.1.2 Methods & Implementation

2.1.3 Results & Validation

2.1.4 Discussion

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

Publication relative to this chapter: [11]

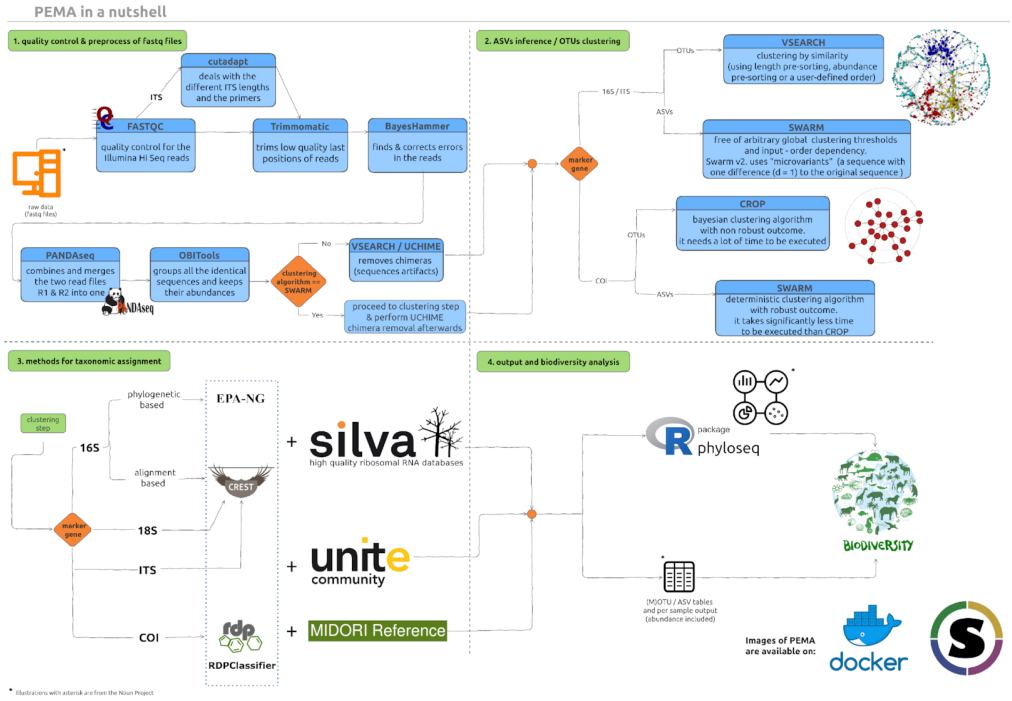

FIGURE 2.1: The PEMA workflow: figure from publication

2.2.1 Introduction

2.2.2 Methods & Implementation

2.2.3 Results & Validation

2.2.4 Discussion

2.3 A workflow for marine Genomic Observatories data analysis

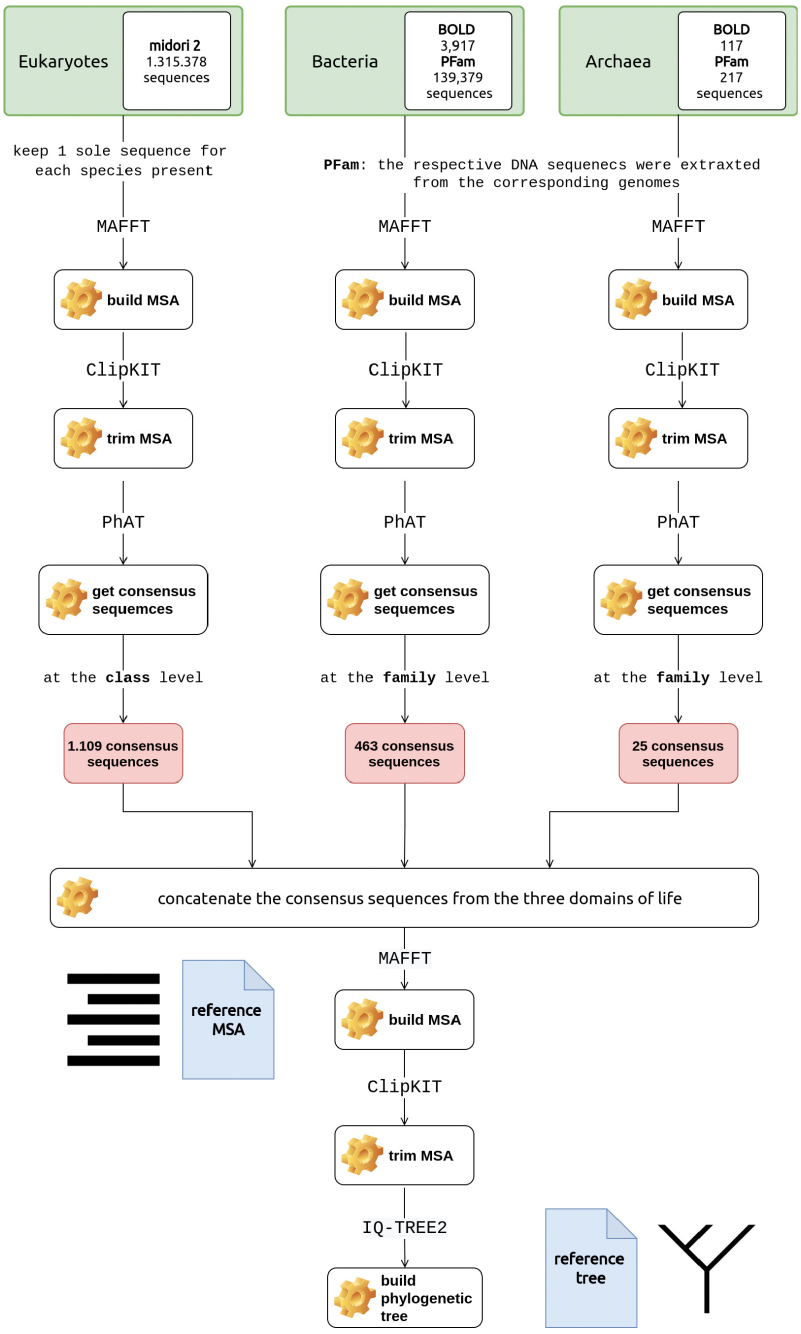

FIGURE 2.2: DARN methodology: figure in the publication

Chapter 3

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

3.1 PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types

Publication relative to this chapter: under submission

3.1.1 Introduction

3.1.2 Methods & Implementation

3.1.3 Results & Validation

3.1.4 Discussion

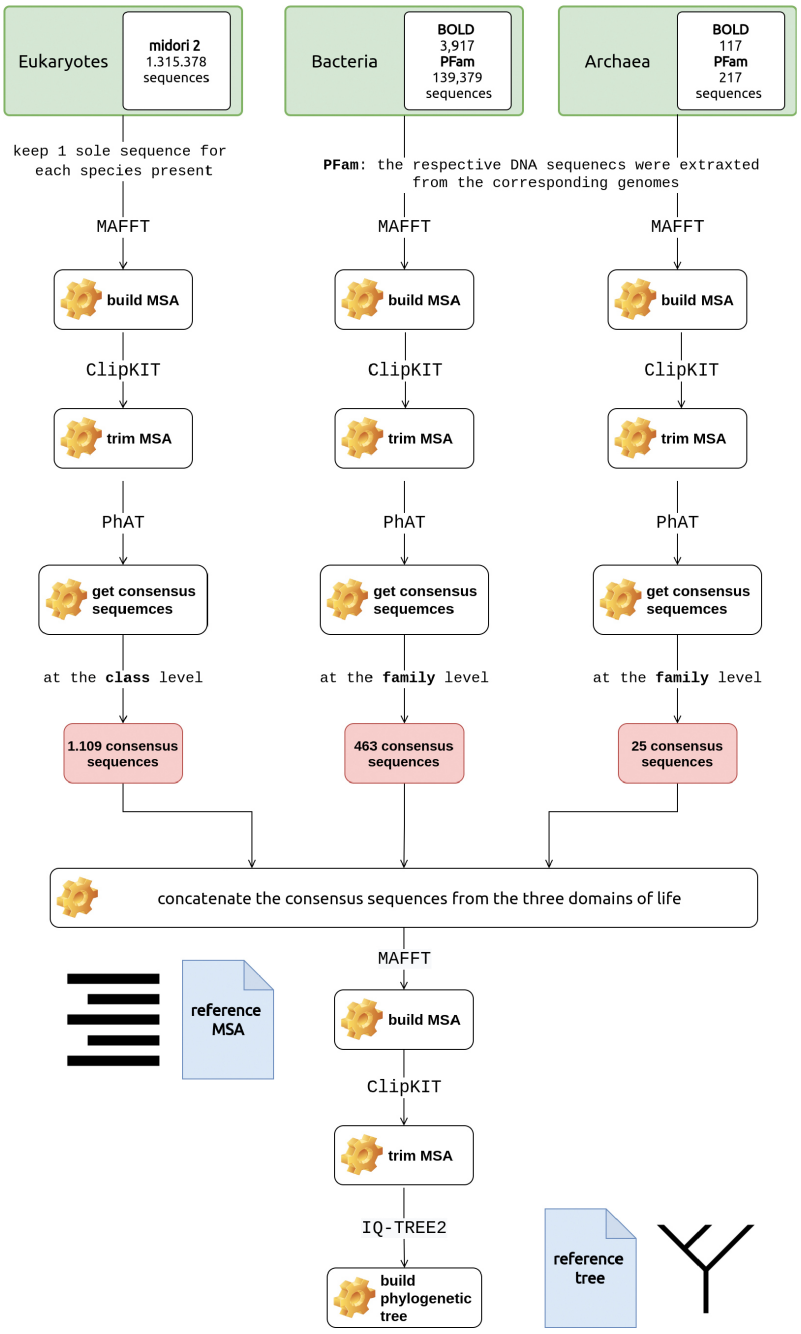

FIGURE 3.1: DARN methodology: figure in the publication

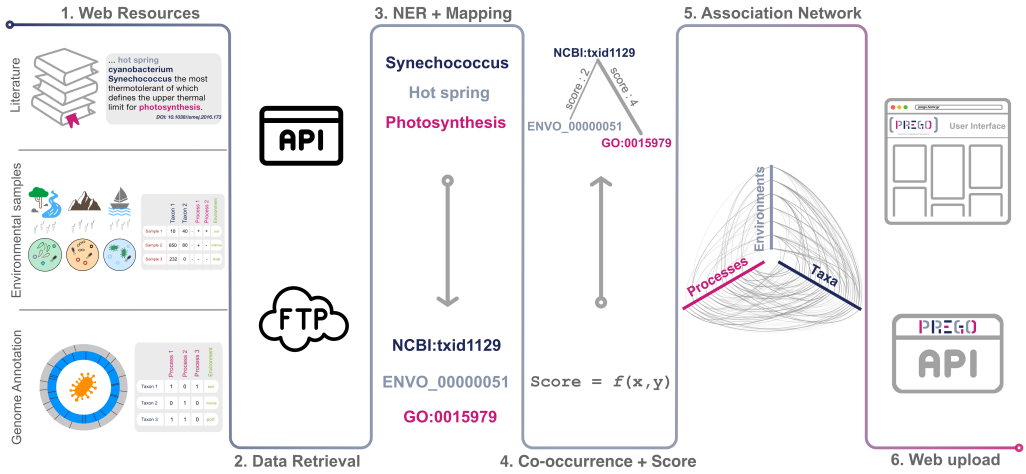

FIGURE 3.2: PREGO methodology: figure in the publication under submission

Chapter 4

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

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

Publication relative to this chapter: [12]

4.1.1 Introduction

Systems Biology is a fundamental field and paradigm that represents a crucial era in Biology. Its functionality and usefulness rely on metabolic networks that model the reactions occurring inside an organism and provide the means to understand the underlying mechanisms that govern biological systems. We address the problem of sampling uniformly steady states of a metabolic network. We use a convex polytope to represent this set. However, the polytopes that result from biological data are of very high dimension (in the order of thousands) and in most, if not all, the cases are considerably skinny. Therefore, to perform uniform 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 from convex polytopes that result from metabolic networks. Its backbone is a Multiphase Monte Carlo Sampling (MMCS) algorithm. We demonstrate the efficiency of our approach by performing extensive experiments on various metabolic networks. Notably, sampling on the most complicated human metabolic network accessible today, Recon3D, corresponding to a polytope of dimension 5 335, took less than 30 hours. To the best of our knowledge, that is out of reach for existing software.

But why being interested in such a task ? The genome of most bacteria are rather short to have issues like that.

However, MAGs can be brought together and build the metabolic model of a whole community!

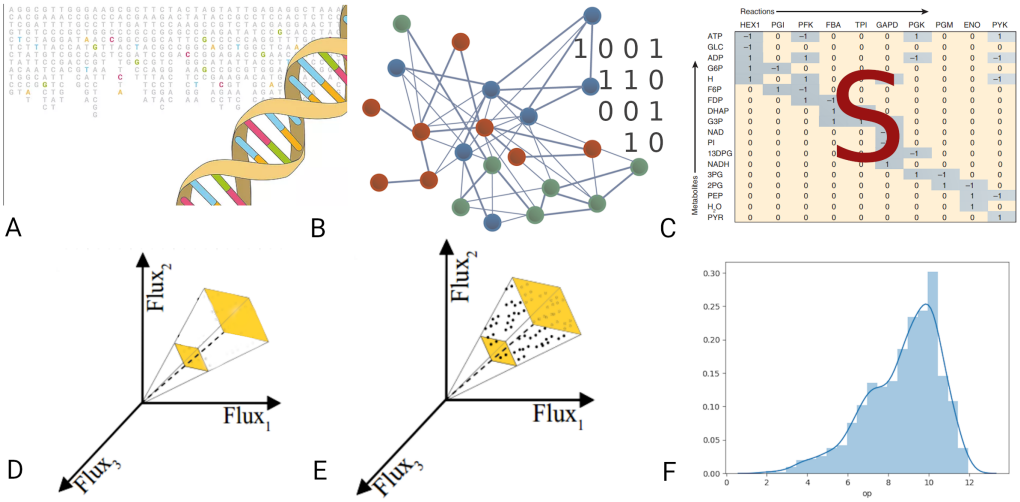

FIGURE 4.1: From DNA sequences to distributions of metabolic fluxes. (A) The genes of an organism provide us with the enzymes that it can potentially produce. Enzymes are like a blueprint for the reactions they can catalyze. (B) Using the enzymes we identify the reactions in the organism. (C) We construct 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.

4.1.2 Methods

Efficient Billiard walk

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

$$a_j^T(p_0 + t_j u_0) = b_j \Rightarrow t_j = (b_j - a_j^T p_0)/a_j^T u_0,\ j \in [k], \tag{4.1}$$

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

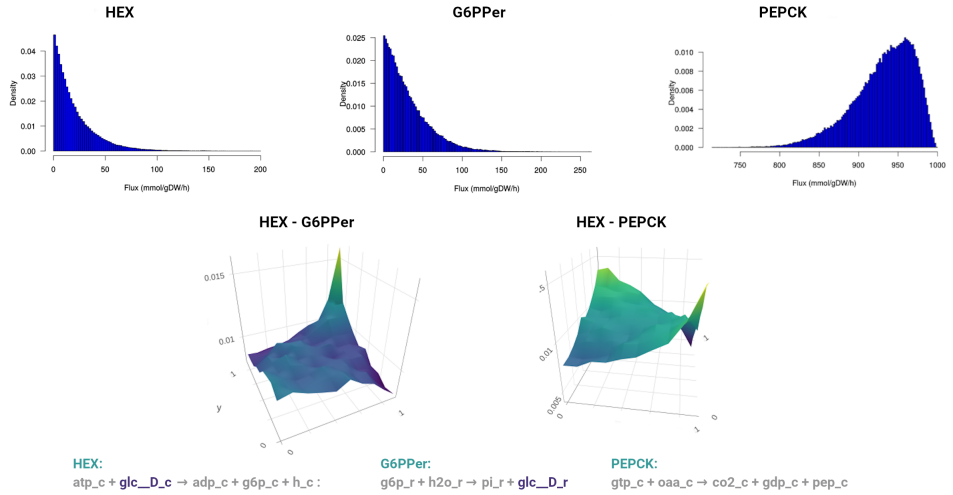

FIGURE 4.2: Flux distributions in the most recent human metabolic network Recon3D [4]. We estimate the flux distributions of the reactions catalyzed by the enzymes Hexokinase (D-Glucose:ATP) (HEX), Glucose-6-Phosphate Phosphatase, Edoplasmic Reticular (G6PPer) and Phosphoenolpyruvate carboxykinase (GTP) (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.

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 4.1.2.

Multiphase Monte Carlo Sampling algorithm

4.1.3 Results

4.1.4 Discussion

Flux sampling at the community level!

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**
$\ell \leftarrow \{p_i + t u_i, 0 \leq t \leq L\}$ { *this is a segment*}
if $\partial P \cap \ell = \emptyset$ **then**
$p_{i+1} \leftarrow p_i + L u_i$ **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

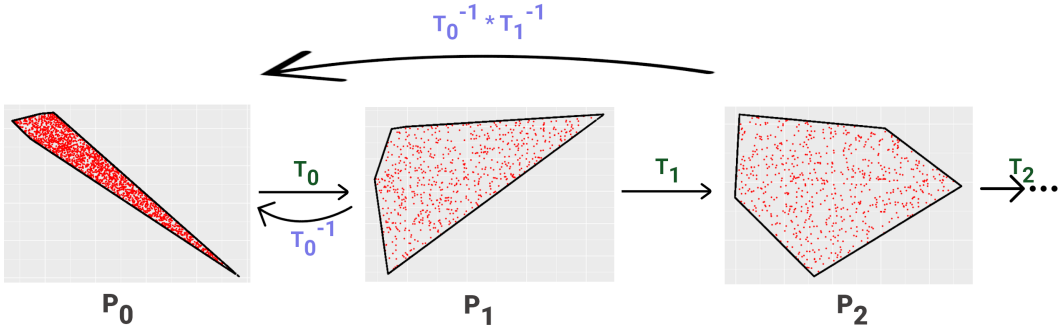

FIGURE 4.3: An illustration of our Multiphase Monte Carlo Sampling algorithm. The method is given an integer n and starts at phase $i = 0$ sampling from P_0. In each phase it samples a maximum number of points λ. If the sum of Effective Sample Size in each phase becomes larger than n before the total 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.

Chapter 5

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

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

5.1 Amplicon & shotgun metagenomic analysis

darn and PEMA will be used at this point, among other software

5.2 Inferring microbial interactions

PREGO and dingo will be used to this end

Chapter 6

0s and 1s in marine molecular research

Publication relative to this chapter: [13]

6.1 Computing resources: a prerequisite & a limitation in modern microbial ecology

6.2 High Performance Computing and Cloudification: scaling up bioinformatics analysis

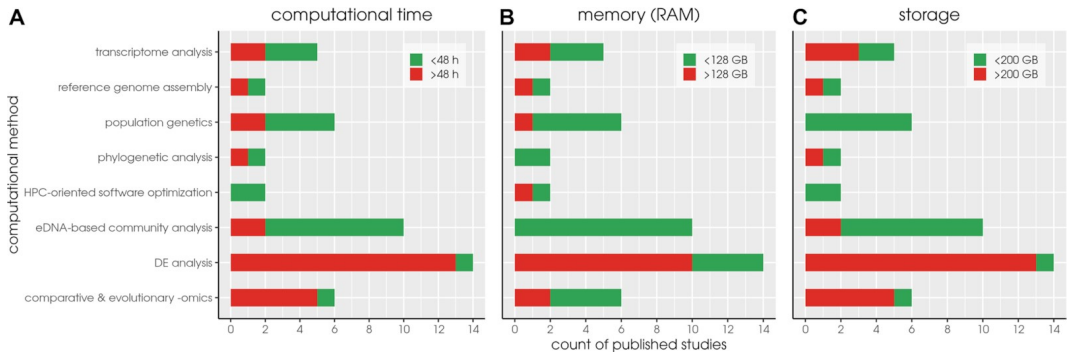

FIGURE 6.1: Computing requirements of the published studies performed on the IMBBC HPC facility over the last decade. Figure from publication.

Chapter 7

Conclusions

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

Appendices

Bibliography

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] H. Zafeiropoulos, L. Gargan, S. Hintikka, C. Pavloudi, and J. Carlsson, "The dark matter investigator (darn) tool: getting to know the known unknowns in coi amplicon data," *Metabarcoding and Metagenomics*, vol. 5, p. e69657, 2021.

[12] A. Chalkis, V. Fisikopoulos, E. Tsigaridas, and H. Zafeiropoulos, "Geometric Algorithms for Sampling the Flux Space of Metabolic Networks," in *37th International Symposium on Computational Geometry (SoCG 2021)* (K. Buchin and E. Colin de Verdière, eds.), vol. 189 of *Leibniz International Proceedings in Informatics (LIPIcs)*, (Dagstuhl, Germany), pp. 21:1–21:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.

[13] 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.