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

developing computational approaches to better understand microbial assemblages

Haris Zafeiropoulos
PhD candidate

Microbial ecology & biogeochemical cycles

a corner-stone for life on earth

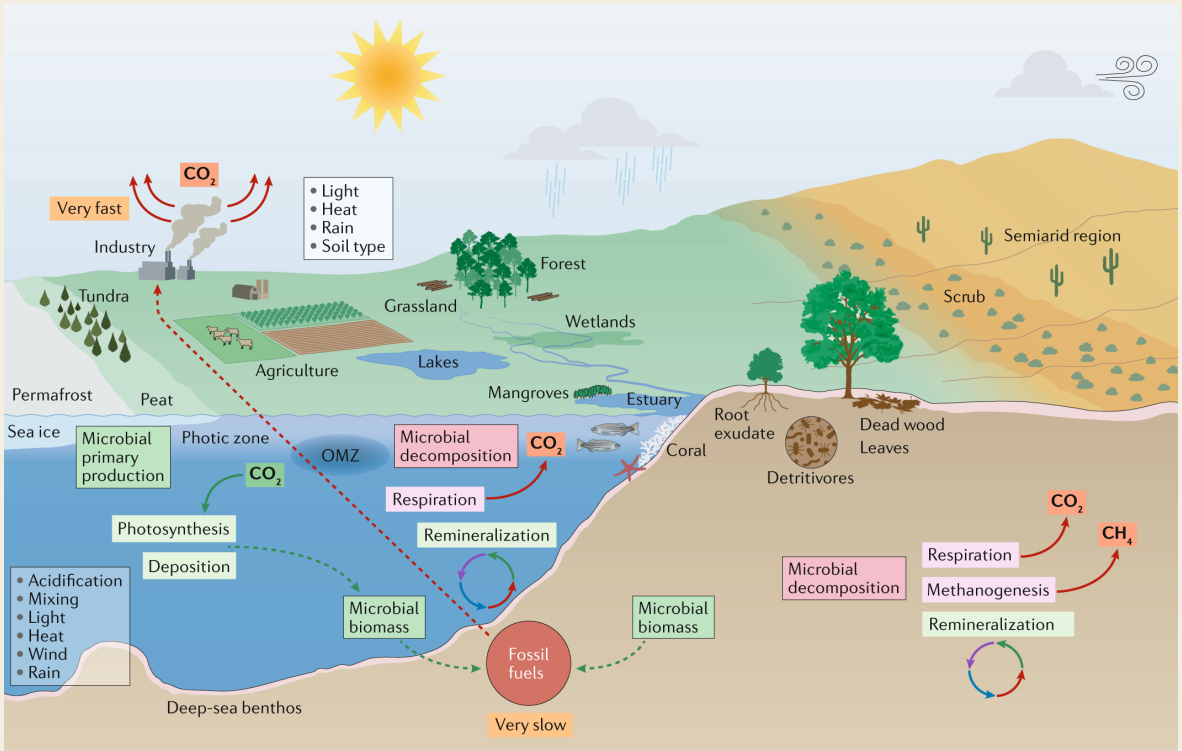


Figure from: Nature Reviews Microbiology 17.9 (2019): 569-586.

- composition
- functions
- interactions

→ power biogeo-chemical cycling

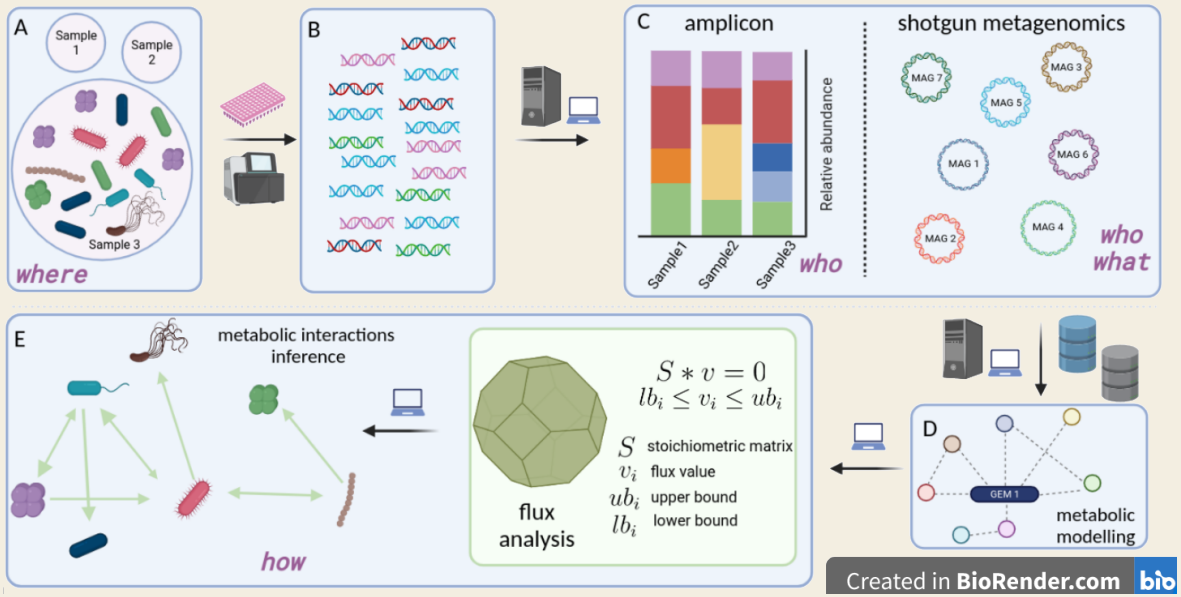
Main questions regarding a microbial community

for a deeper understanding of such assemblages

Community structure ***who***	Functional potential ***what***	Microbial interactions ***why / how***
eveyrone is everywhere	*zero-sum game*	*the entagnled bank*

Reverse ecology

transforming ecology into a high-throughput field



community ecology studies with no *a priori* assumptions about the organisms under consideration by exploiting advances in systems biology and genomic metabolic modeling

High Throughput technologies

a new era bringing its own challenges

- biology-oriented issues
- techonology-oriented issues
- computing requirements
- multiple channels of information

Aims and objectives

- to enhance the analysis of microbiome data by building algorithms and software that address limitations and on-going computational challenges

- to exploit state-of-the-art methods to identify taxa and functions that play a key part in microbial community assemblages in hypersaline sediments

1. **Chapter 1**: Introduction
2. **Chapter 2**: Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
    2.1 **Chapter 2.1**: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes
    2.2 **Chapter 2.2**: The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4**: A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** 0s and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

eDNA metabarcoding

for biodiversity assessment

Marker genes

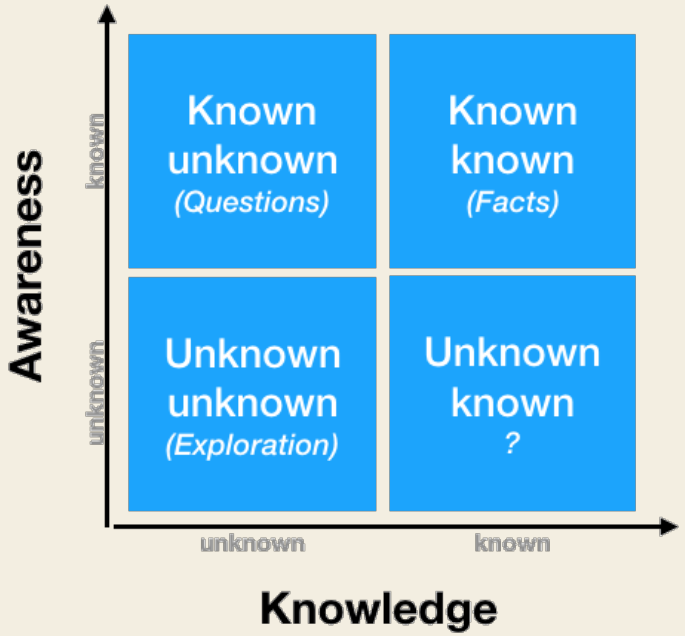
1. **16S rRNA:** Bacteria, Archaea
2. **12S rRNA:** Vertebrates
3. **18S rRNA:** Small eukaryotes, Metazoa
4. **ITS:** Fungi
5. **COI:** Eukaryotes
6. ***rbcl*:** Plants
7. ***dsrb*:** Bacteria, Archaea
8. ...

Methodology

- Sampling
- Extraction
- Bioinformatics
- Biodiversity analysis

Bioinformatics challenges

for the analysis and the interpretation of amplicon data



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

Aim of the study and contribution

To build an open source pipeline that bundles state-of-the-art bioinformatics tools for all necessary steps of amplicon analysis and aims to address:

- one-stop-shop for several marker genes & approaches
- easey-to-set & easy-to-use
- scalable
- flexible

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

Aim of the study and contribution

To build a framework for extracting 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).

We argue that the vast majority of these sequences represent microbial taxa, such as bacteria and archaea

1. **Chapter 1**: Introduction
2. **Chapter 2**: Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
    2.1 **Chapter 2.1**: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes
    2.2 **Chapter 2.2**: The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4**: A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** 0s and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

Chapter 3: PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types

Aim of the study and contribution

To build a hypothesis generation resource based on associations between:

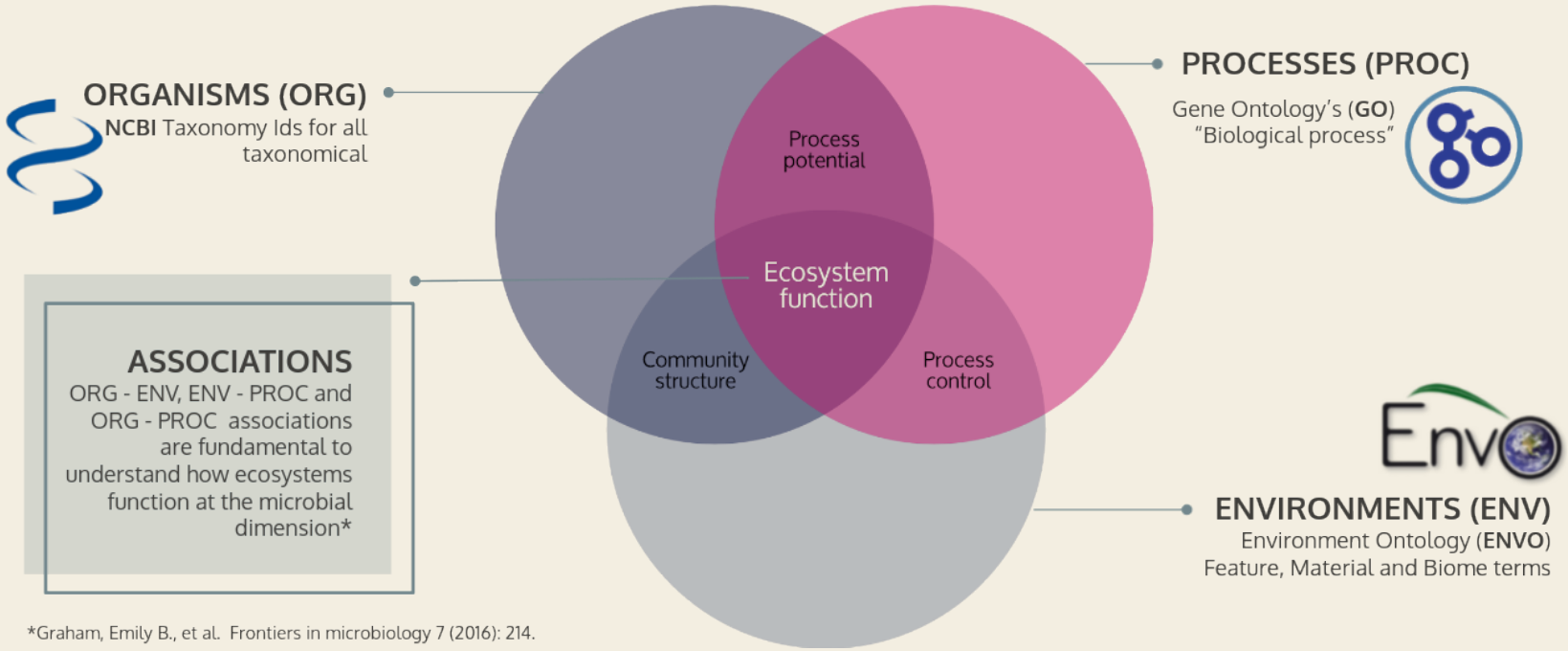
- *organisms* and the *environments* they inhabit
- *organisms* and the biological *processes* they are involved with
- *processes* and the *environments* where they occur

To this end, associations among such terms were exported from:

- the publicaly available literature
- genome and omics' studies, their results and their corresponding metadata

PREGO and its referring terms

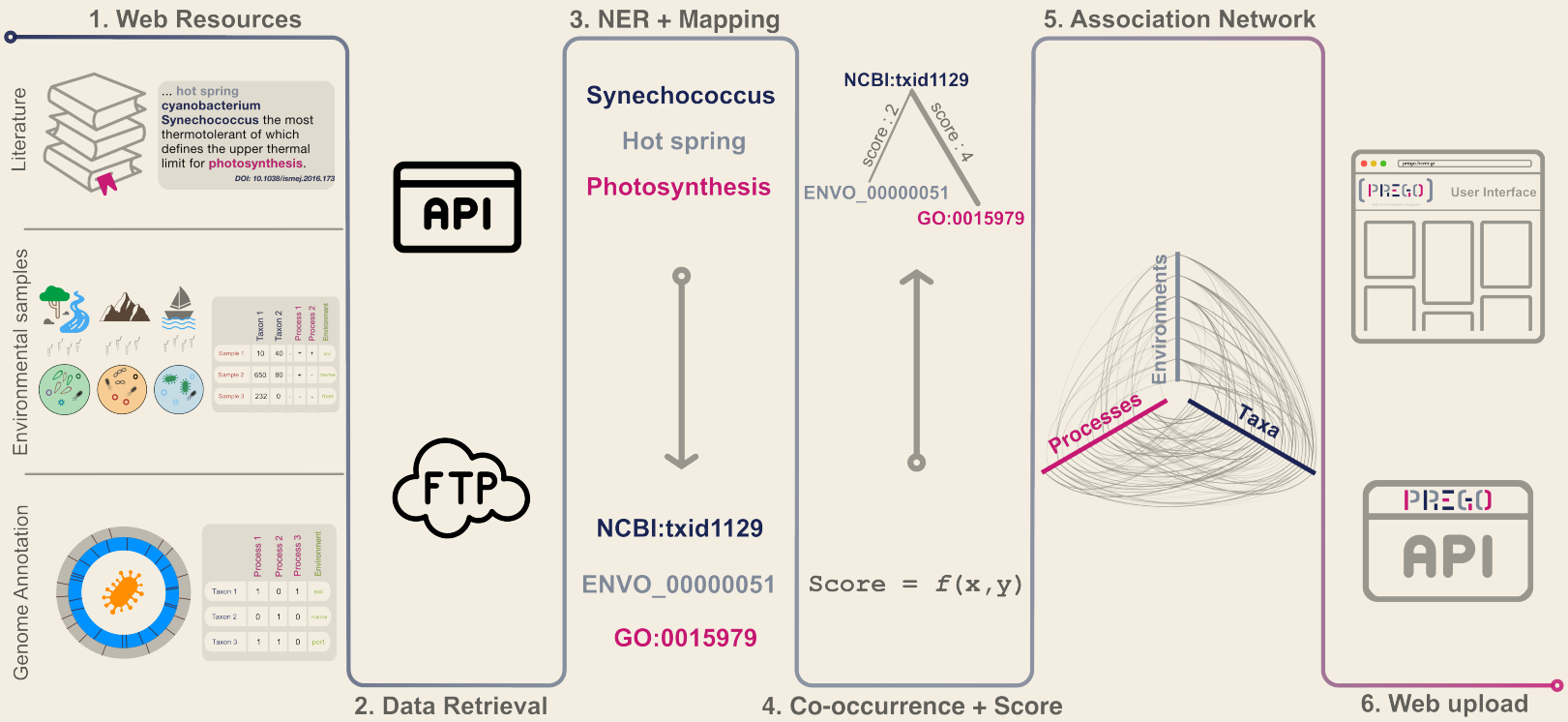
the fundamental role of ontologies



*Graham, Emily B., et al. Frontiers in microbiology 7 (2016): 214.

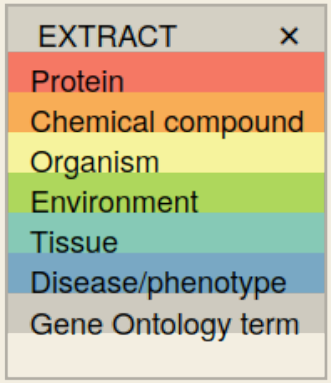
Methods / Implementation

3 channels of information - 1 framework



Named Entity Recognition & the literature channel

exporting associations from publicly available publications



Identification of potentially important pathways missing from the model

From the metagenomic bins, we were able to identify two metabolic processes that were not previously included in the model. A number of MAGs (bin.59, bin.15, bin.73) clustered to the KEGG genomes of freshwater sulfur-oxidizing autotrophs capable of denitrification, *Sulfuritalea hydrogenivorans* [41], and *Sulfuricella denitrificans* [42]. These MAGs contained the diagnostic genes for carbon fixation (*rbcLS*), sulfur cycling (*dsrAB*), and denitrification (*nosZ*). One MAG (bin.59) also clustered with iron oxidizing autotroph *Sideroxydans lithotrophicus* ES-1. Bin.59 is the most relatively abundant bin from 17 to 21 m depth. Thus, if this MAG is associated with iron oxidation, it also contains sulfur-cycling genes that add to metabolic flexibility, which was previously observed [40]. The model did not include sulfide oxidation with nitrate, so it is unclear from the current model predictions where this process is expected to occur within the water column to compare to the MAG distributions.

Example text from Arora-Williams et al. Microbiome 6.1 (2018): 1-16.

The *Environmental Samples* and the *Annotated Genomes and Isolates* channels

the role of metadata

Sample metadata [-]

Collection date:	11/1/11
Elevation:	200
Environment (biome):	soil
Environment (feature):	nosZ
Environment (material):	soil DNA
Environmental package:	MIGS/MIMS/MIMARKS.soil
Geographic location (depth):	15-20cm
Instrument model:	454 GS FLX Titanium
Investigation type:	miens-survey
NCBI sample classification:	410658
Project name:	EcoFINDERS

Project Information	
Cultured	No
Ecosystem	Environmental
Ecosystem Category	Aquatic
Ecosystem Subtype	Oceanic
Ecosystem Type	Marine

MG-RAST ID	name	biome	feature	material	sample	library	location	country	coordinates	download
mgm4702467.3	06032015b_S2_L001_R2_001	Large lake biome	lake	water	mgs485560	mgl485562	Cincinnati	USA	39.11, -84.5	metadata submitted
mgm4702469.3	06052015a_S3_L001_R2_001	Large lake biome	lake	water	mgs485566	mgl485568	Cincinnati	USA	39.11, -84.5	metadata submitted
mgm4702471.3	06032015a_S1_L001_R2_001	Large lake biome	lake	water	mgs485554	mgl485556	Cincinnati	USA	39.11, -84.5	metadata submitted

Co-mentioning and scoring scheme

which are the most worthy and relevant associations

- → genome annotation oriented associations: fixed scores
- → associations in the *Environmental Samples* channel are scored based on the number of samples they co-occur.
- → similarly, in the *Literature* channel, based on the number of publications

		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	$c_{.,y}$	$c_{.,0}$	$c_{.,.}$

Environmental samples score:

$$score_{x,y} = 2.0 * \sqrt{\frac{c_{x,y}}{c_{.,y}}}^{a} \quad (1)$$

Associations between entities of PREGO

after metadata retrieveal and co-occurrence analysis

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

A real case hypothesis generation scenario

Posidonia and its microbiome

What about ***Posidonia*** ?
Literature suggests that Planctomycetes
and especially *Blastopirellula*
and *Rhodopirellula* are commonly
found in its microbiome.
Why so ?
You may have a look here

Conclusions

on PREGO and its associations

- Literature
- similar number of molecular functions in all cases indicates the robustness of the main metabolic processes required for life
- , 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

1. **Chapter 1**: Introduction
2. **Chapter 2**: Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
    2.1 **Chapter 2.1**: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes
    2.2 **Chapter 2.2**: The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4**: A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** 0s and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

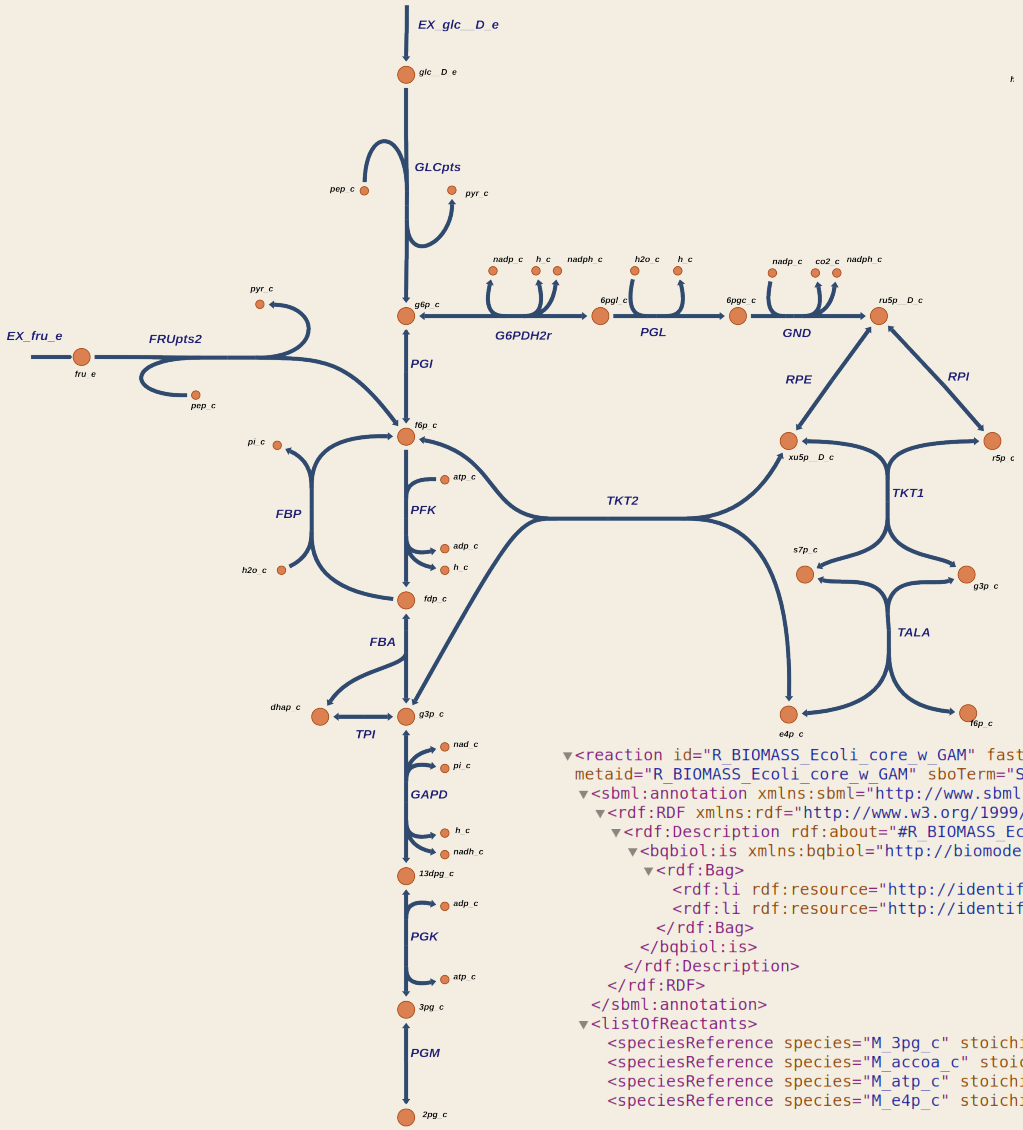
Chapter 4: A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks

Aim of the study and contribution

Flux sampling is a computationally intensive task, especially as the dimension of the polytopes derived from the metabolic model under study increases.
Microbial genome-scale models correspond to relatively low dimensional polytopes, but that is not the case for models integrating multiple GEMs. Further, more often than not, metabolic models of a host and a microbe is Therefore, to allow flux sampling at high dimensional polytopes we introduce a Multi-phase Monte Carlo Sampling (MMCS) algorithm.

Metabolic modelling

and the biomass function




```
▼<reaction id="R_BIOMASS_Ecoli_core_w_GAM" fast="false" reversible="false" name="Biomass Objective Function with GAM"
 metaid="R_BIOMASS_Ecoli_core_w_GAM" sboTerm="SBO:0000629" fbc:upperFluxBound="cobra_default_ub" fbc:lowerFluxBound="cobra_0_bound">
 ▼<sbml:annotation xmlns:sbml="http://www.sbml.org/sbml/level3/version1/core">
  ▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   ▼<rdf:Description rdf:about="#R_BIOMASS_Ecoli_core_w_GAM">
    ▼<bqbiol:is xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
     ▼<rdf:Bag>
        <rdf:li rdf:resource="http://identifiers.org/bigg.reaction/BIOMASS_Ecoli_core_w_GAM"/>
        <rdf:li rdf:resource="http://identifiers.org/metanetx.reaction/MNXR96280"/>
      </rdf:Bag>
     </bqbiol:is>
    </rdf:Description>
   </rdf:RDF>
  </sbml:annotation>
 ▼<listOfReactants>
    <speciesReference species="M_3pg_c" stoichiometry="1.496" constant="true"/>
    <speciesReference species="M_accoa_c" stoichiometry="3.7478" constant="true"/>
    <speciesReference species="M_atp_c" stoichiometry="59.81" constant="true"/>
    <speciesReference species="M_e4p_c" stoichiometry="0.361" constant="true"/>
```

Metabolic models allow us

to move from a metabolic map

to mathematical structures

the study of which may provide

fundamental biological insight

1. **Chapter 1**: Introduction
2. **Chapter 2**: Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
    - 2.1 **Chapter 2.1**: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes
    - 2.2 **Chapter 2.2**: The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4**: A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** 0s and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

Chapter 5: Deciphering the functional potential of a hypersaline marsh microbial mat community

Aim of the study and contribution

1. **Chapter 1**: Introduction
2. **Chapter 2**: Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
    2.1 **Chapter 2.1**: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes
    2.2 **Chapter 2.2**: The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4**: A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** 0s and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

Chapter 6: 0s and 1s in marine molecular research: a regional HPC perspective

Aim of the study and contribution

1. **Chapter 1**: Introduction
2. **Chapter 2**: Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
    2.1 **Chapter 2.1**: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes
    2.2 **Chapter 2.2**: The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4**: A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** 0s and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

Chapter 7: Conclusions

- Bioinformatics approaches enhance microbial diversity assesment based on HTS data
- Containerization technologies and e-infrastructures provide the means for computational capacity and reproducibility
- High quality metadata enable efficient exploitation of sequecning data in a meta-analysis level

Chapter 7: Conclusions

and future perspectives

- Markov Chain Monte Carlo approaches enable flux sampling in high-dimensional polytopes

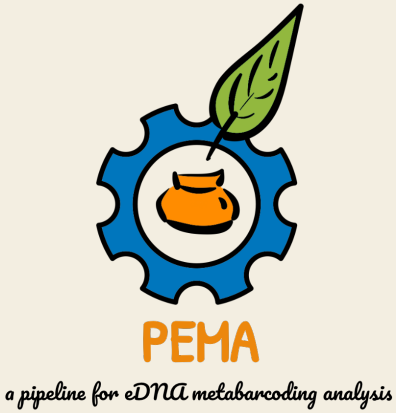
- Hypersaline mats host a great range of novel taxa & their functioning might be subject to anaplerotic reactions

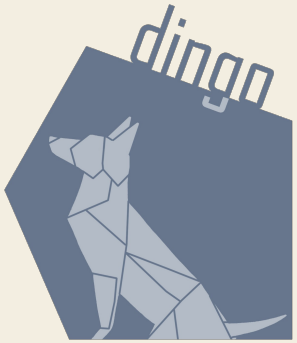
Perspective for future work

"a combination of quantitative high - throughput experiments and predictive metabolic models can elucidate the genotype - phenotype map of microbial metabolic strategies" providing great insight on the evolvability of metabolic decisions and on how such decisions affect microbial coexistence in the communities.

Wrap-up

software tools





github.com/hariszaf/pema

github.com/hariszaf/darn

github.com/lab42open-team/ the `prego*` repositories

github.com/GeomScale/dingo

Publications

[1] **Zafeiropoulos, H.**, Paragkamian, S., Ninidakis, S., Pavlopoulos, G.A., Jensen, L.J. & Pafilis, E. (2022). PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types. Microorganisms 10(2), 293.

[2] **Zafeiropoulos, H.**, Gargan, L., Hintikka, S., Pavloudi, C. & Carlsson, J. (2021). The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data. Metabarcoding and Metagenomics, 5, e69657.

[3] Chalkis, A., Fisikopoulos, V., Tsigaridas, E. & **Zafeiropoulos, H.** (2021). Geometric algorithms for sampling the flux space of metabolic networks, 37th International Symposium on Computational Geometry (SoCG 2021).

[4] **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.

[5] Polymenakou, P.N., Nomikou, P., **Zafeiropoulos, H.**, ..., Kyrpides, N.C., Kotoulas, G. & Magoulas, A. (2021). The santorini volcanic complex as a valuable source of enzymes for bioenergy. Energies, 14(5), p.1414.

[6] **Zafeiropoulos, H.**, Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C. & 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), giaa022.

[7] Pavloudi, C. & **Zafeiropoulos, H.** (2022) Deciphering the community structure and the functional potential of a hypersaline marsh microbial mat community ***(under review at FEMS Microbiology Ecology)***

[8] Garza, D.R., Gonze, D., **Zafeiropoulos, H.**, Liu, B. & Faust, K., (2022) Metabolic models of human gut microbiota: advances and challenges ***(under review at Cell systems)***

[9] Paragkamian, S., Sarafidou, G., ..., **Zafeiropoulos, H.**, Arvanitidis, C., Pafilis, E. & Gerovasileiou, V. Automating the curation process of historical literature on marine biodiversity using text mining: the DECO workflow ***(under review in Frontiers in Marine Science)***

Acknowledgments

funding & grants

GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION

H.F.R.I.
Hellenic Foundation for Research & Innovation

Google
Summer of Code

Acknowledgments

people and more

My promotors:

Dr. Pafilis E.

Prof. Nikolaou Chr.

Prof. Ladoukakis

the rest of my 7-member committee:

Prof. Dina Lika

Prof. Panagiotis Sarris

Prof. Jens Carlsson

Prof. Karoline Faust

Special thanks to:

PhD Paragkamian S.

Dr. Gargan L.

Dr. Hintikka S.

Mr. Ninidakis St.

Mr. Potirakis Ant.

Dr. Chalkis A.

Dr. Fisikopoulos V.

Prof. Tsigaridas E.

My mojo:

Dr. Pavloudi C.

My corner:

Would not be here

if it was not with you.