

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



1. Microbial ecology: a short introduction
2. Bioinformatics methods for microbial diversity assessment
 - 2.1 pema: a metabarcoding pipeline
 - 2.2 darn: known unknowns in COI amplicon data
3. PREGO: a knowledge-base for organisms - environments - processes associations
4. dingo: a Python library for metabolic flux sampling
5. Os & 1s in molecular biology
6. microbetag: enhancing microbial interactions inference from co-occurrence networks
7. Publications

Microbial ecology & biogeochemical cycles

a corner-stone for life on earth

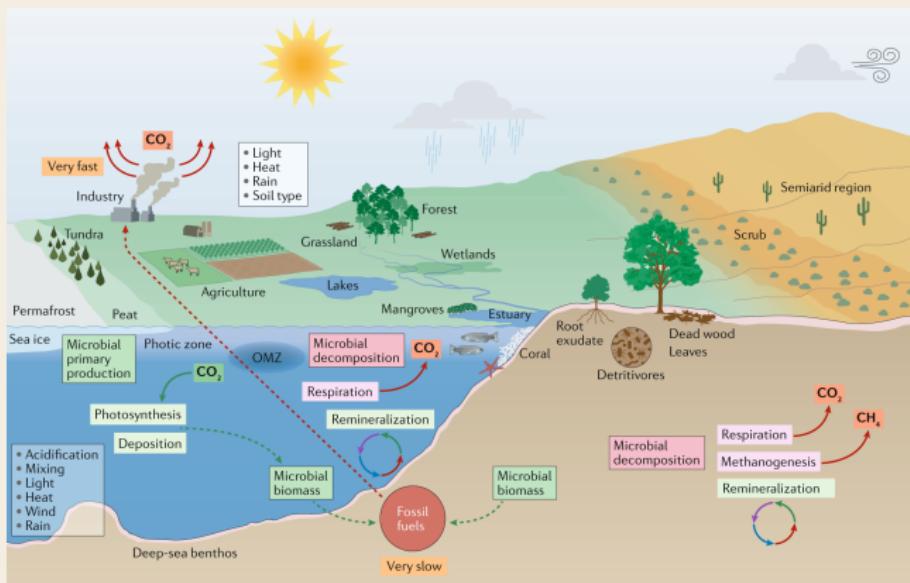


Figure from: Cavicchioli et al. Nature Reviews Microbiology 17.9 (2019): 569-586.

Questions to address

for a deeper understanding of microbial assemblages

Community
structure

who

everyone is everywhere

Functional
potential

what

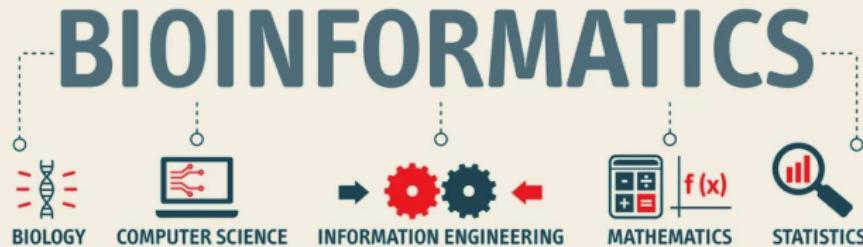
zero-sum game

Microbial
interactions

why

the entangled bank

We are living in a computational era
both a challenge & an opportunity



1. Microbial ecology: a short introduction
2. Bioinformatics methods for microbial diversity assessment
 - 2.1 pema: a metabarcoding pipeline
 - 2.2 darn: known unknowns in COI amplicon data
3. PREGO: a knowledge-base for organisms - environments - processes associations
4. dingo: a Python library for metabolic flux sampling
5. Os & 1s in molecular biology
6. microbetag: enhancing microbial interactions inference from co-occurrence networks
7. Publications

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. ...

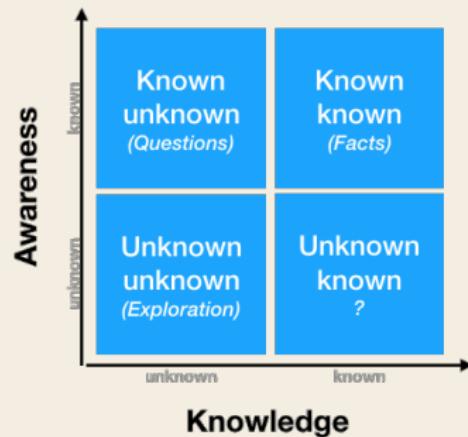
Methodology



- Sampling
- Extraction
- Bioinformatics
- Biodiversity analysis

Bioinformatics challenges

for the analysis and the interpretation of amplicon data





PEMA

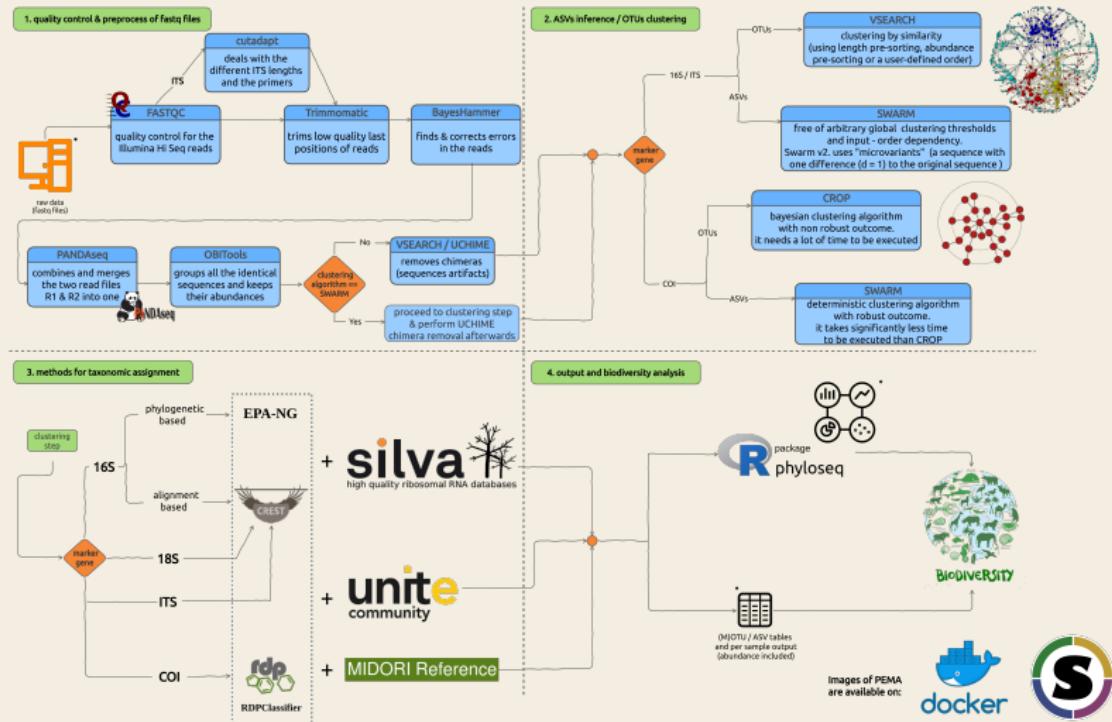
a pipeline for eDNA metabarcoding analysis

<https://github.com/hariszaf/pema>

PEMA features

one step at a time!

PEMA in a nutshell

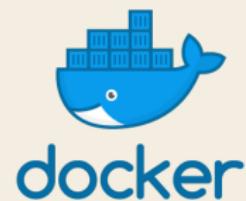


PEMA coding insights

Being a geek just for a bit !

```
for(int i : range(1,  
    in := "in_$i.tx  
    sys date > $in  
  
    out := "out_$i.t  
    task( out <- in  
        sys echo Tas  
    }  
}
```

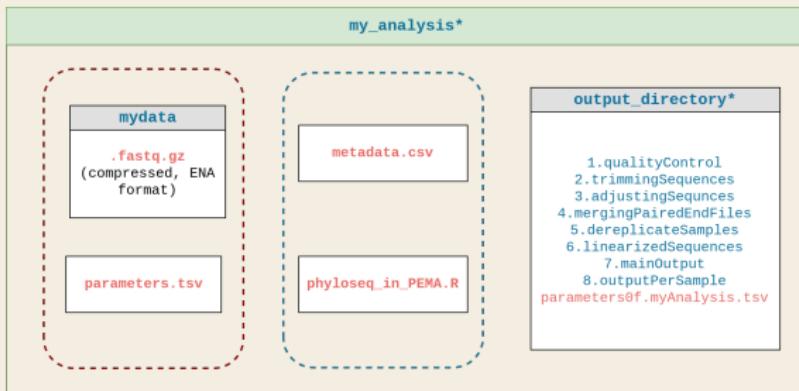
BigDataScript
programming language



Containerization

Mount your I/O

give & take

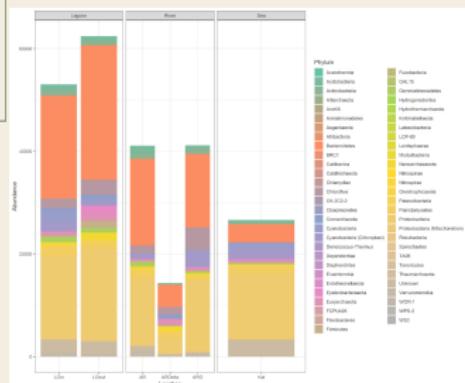


text file directory

*user can edit the name
(the rest **need** or will have the exact names shown)

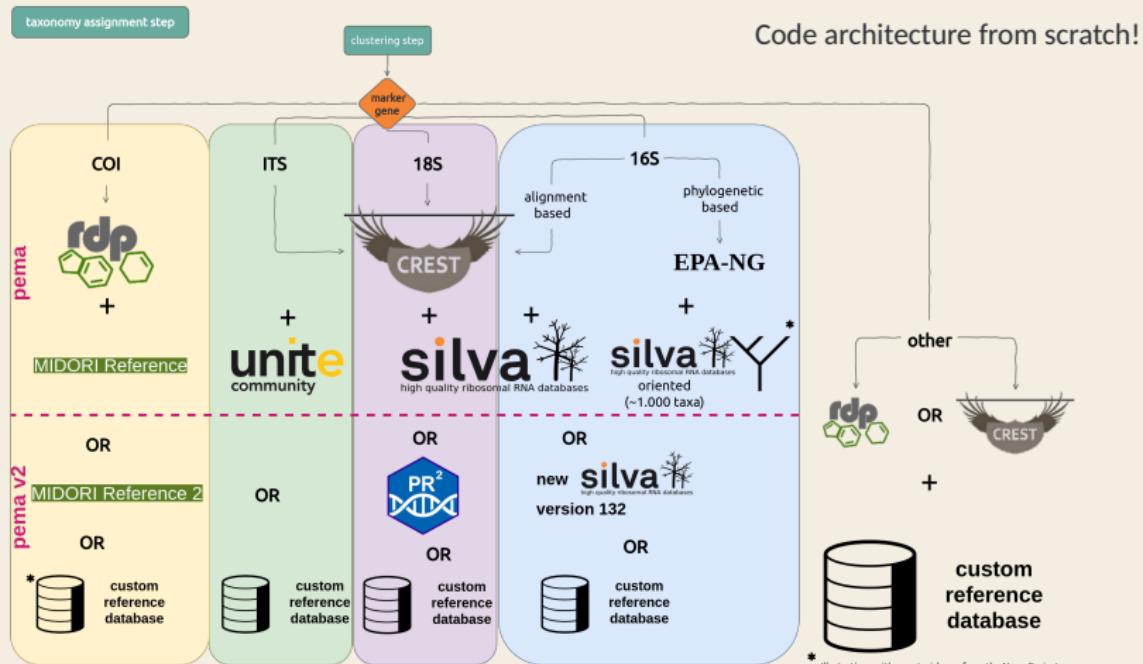
— mandatory input files
— optional input files

	Sample 1	Sample 2	Sample 3	Sample 4
Taxon 1	1	0	1	2
Taxon 2	0	1	0	2
Taxon 3	1	1	0	4



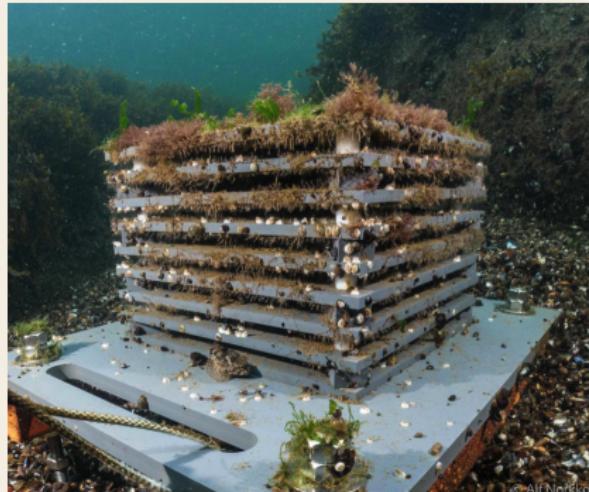
PEMA v.2

addressing some of the challenges



* Illustrations with an asterisk are from the Noun Project

Latest PEMa version *addressing the challenges of the community*



ASSEMBLE 
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

MBON
Marine Biodiversity
Observation Network

pema:v.2.1.4 includes:

1. analysis of 12S rRNA data now supported ([12S Vertebrate Classifier v2.0.0-ref database](#))
2. PR2 as an alternative reference database for the case of 18S rRNA
3. the ncbi-taxonomist tool was added to return the NCBI Taxonomy Id of the taxonomies found

How to and further documentation at pema.hcmr.gr

A place for sharing news as well as thoughts and remarks on how to run metabarcoding analyses using the PEMA containers or other software. Made by Haris Zafeiropoulos.

- Home
- Get PEMA
- Basics for running PEMA
- PEMA output
- Running on HPC
- Running on personal computer
- 16/18S analysis
- COI analysis
- ITS analysis
- Training the CREST classifier
- Training the RDPClassifier
- Tuning tuning tuning!
- GitHub repo
- Blog

© 2021. All rights reserved.

PEMA investigating metabarcoding

Welcome

Hey there!

This is the PEMA main site for *how to use* and further metabarcoding tips and hints!

You may find PEMA as a Docker and as a Singularity container.

Here is the [PEMA GitHub repository](#) if you want to have a look on the source code and why not, to contribute to!

For any running issues you may have, or for any further features you would like to see included on PEMA, you can reach us through the [PEMA Gitter community](#) or at pema@hcmr.gr.

Thanks for your interest on PEMA! Keep metabarcoding!

The PEMA team



PEMA
a pipeline for eDNA metabarcoding analysis



darn

<https://github.com/hariszaf/darn>

Dark mAtteR iNvesigator

investigating known unknown
in COI amplicon data

What is all the unassigned OTUs/ASVs
in COI amplicon data?

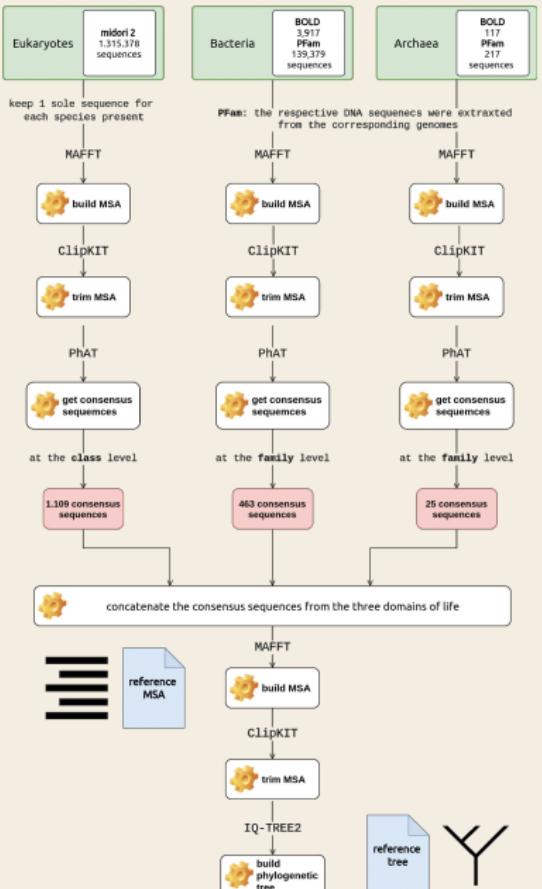
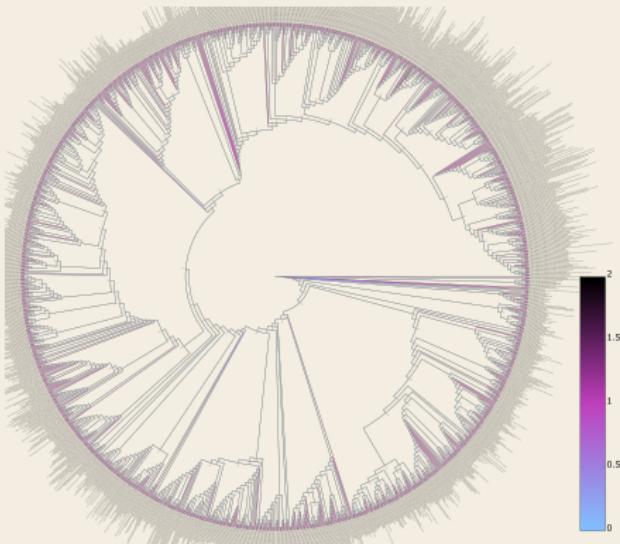


Figure from: [10.3897/mbmg.5.69657](https://doi.org/10.3897/mbmg.5.69657)

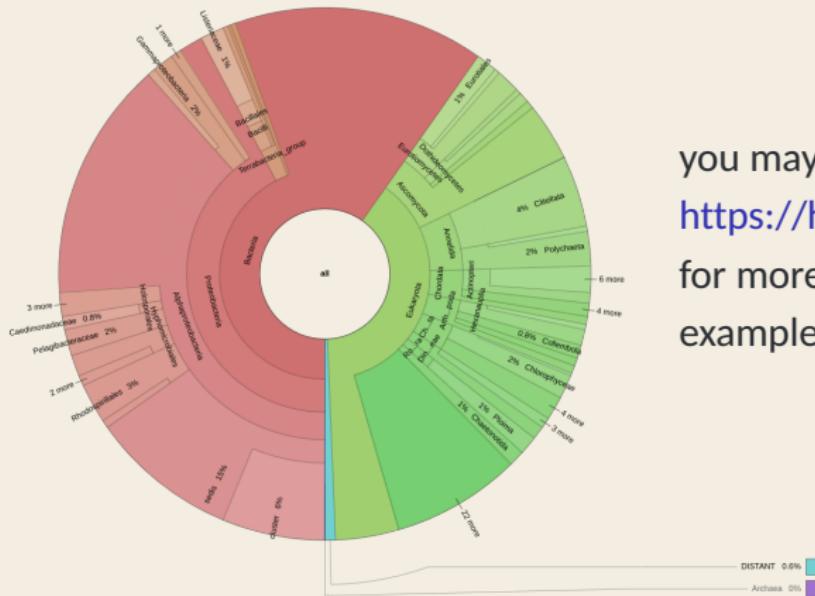
Phylogeny of the COI consensus sequences retrieved *the tree that DARN makes use of*



the consensus sequences have been placed in their corresponding taxonomic branches, proving the tree valid

Bacteria are everywhere!

... Archaea too!



you may have a look at

<https://hariszaf.github.io/darn/>
for more DARN output
example cases

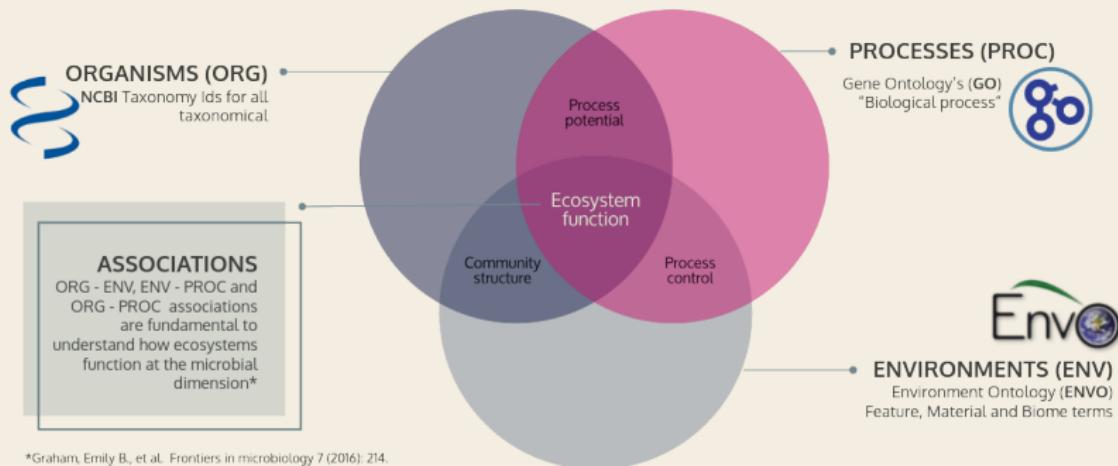
1. Microbial ecology: a short introduction
2. Bioinformatics methods for microbial diversity assessment
 - 2.1 pema: a metabarcoding pipeline
 - 2.2 darn: known unknowns in COI amplicon data
3. PREGO: a knowledge-base for organisms - environments - processes associations
4. dingo: a Python library for metabolic flux sampling
5. Os & 1s in molecular biology
6. microbetag: enhancing microbial interactions inference from co-occurrence networks
7. Publications



PROCESS ENVIRONMENT ORGANISM

related repositories under
<https://github.com/orgs/lab42open-team>

PREGO as processes - environments - organisms and how to link them



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

Metadata example

from various metagenome repositories

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:	metres-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	mg485562	Cincinnati	USA	39.11, -84.5	
mgm4702469.3	06052015a_S3_L001_R2_001	Large lake biome	lake	water	mgs485566	mg485568	Cincinnati	USA	39.11, -84.5	
mgm4702471.3	06032015a_S1_L001_R2_001	Large lake biome	lake	water	mgs485554	mg485556	Cincinnati	USA	39.11, -84.5	

MG-RAST
metagenomics analysis server

Named Entity Recognition

tagging the literature

Identification of potentially important pathways missing from the model

EXTRACT	X
Protein	
Chemical compound	
Organism	
Environment	
Tissue	
Disease/phenotype	
Gene Ontology term	

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 hydrogentivorans* [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.

PREGO methodology

co-occurrence again!

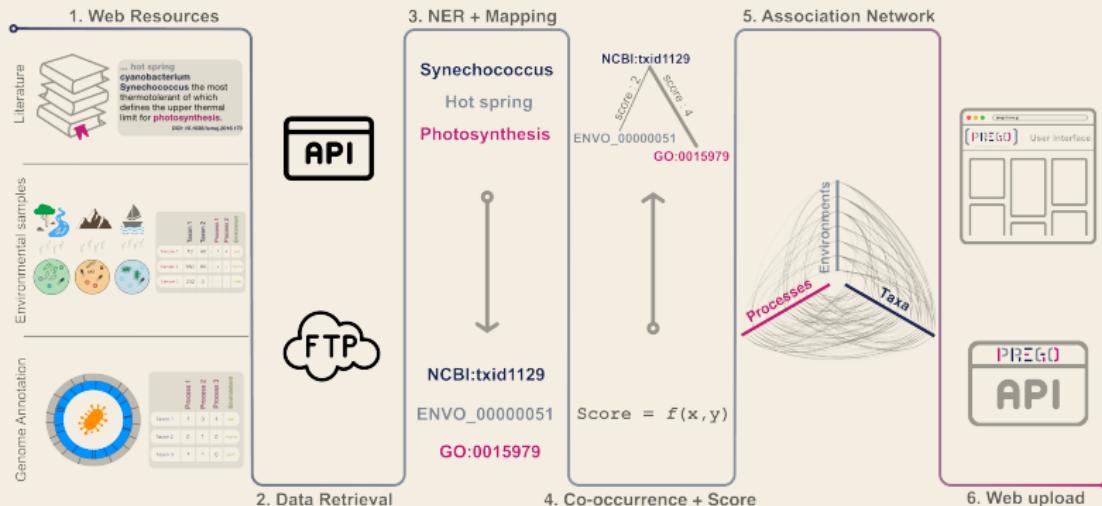
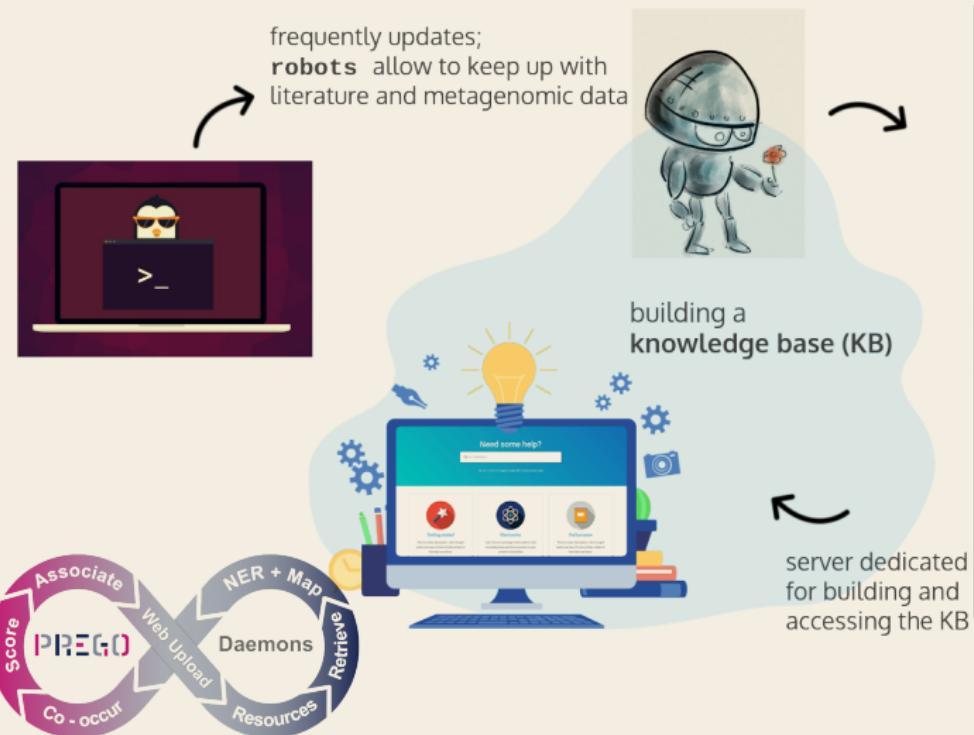


Figure from the PREGO publication that is now under review.

Building a knowledge-base

development and information technology operations



PREGO in action

looking for environments a taxon is present

Desulfatiglans anilini DSM 4660 [1121399]

Synonyms: Desulfatiglans anilini DSM 4660, D. anilini DSM 4660, D anilini DSM 4660, Desulfatiglans DSM 4660, Desulfatiglans str. DSM 4660 ...

Environments Biological Processes Molecular Function Documents Downloads

Literature

Search:

Name	Z-score	Confidence
Oil seep	3.0	★★★★★
Marine mud	3.0	★★★★★
Marine sediment	2.8	★★★★★
Brackish water	2.4	★★★★★
Oil reservoir	2.2	★★★★★
Anaerobic sediment	2.0	★★★★★
Cold seep	1.8	★★★★★
Contaminated sediment	1.7	★★★★★
Neritic sub-litoral zone	1.4	★★★★★
Oil spill	1.4	★★★★★
Petroleum	1.1	★★★★★
Sea floor	1.1	★★★★★

Showing 1 to 12 of 12 entries

Environments Biological Processes Molecular Function Documents Downloads

Literature

Search:

Name	Z-score	Confidence
benzoyl-CoA catabolic process	4.3	★★★★★
Benzene catabolic process	3.6	★★★★★
Acetone metabolic process	3.5	★★★★★
Phenanthrene catabolic process	3.5	★★★★★
Sulfate reduction	3.3	★★★★★
Ketone body catabolic process	3.2	★★★★★
Naphthalene catabolic process	3.2	★★★★★
Alkane catabolic process	3.1	★★★★★
Benzoate catabolic process	3.0	★★★★★
Denitrification pathway	2.8	★★★★★
Sulfide ion homeostasis	2.6	★★★★★
Ketone catabolic process	2.6	★★★★★
Methanogenesis	1.8	★★★★★
Electron transport chain	1.0	★★★★★

Showing 1 to 14 of 14 entries

1. Microbial ecology: a short introduction
2. Bioinformatics methods for microbial diversity assessment
 - 2.1 pema: a metabarcoding pipeline
 - 2.2 darn: known unknowns in COI amplicon data
3. PREGO: a knowledge-base for organisms - environments - processes associations
4. dingo: a Python library for metabolic flux sampling
5. Os & 1s in molecular biology
6. microbetag: enhancing microbial interactions inference from co-occurrence networks
7. Publications



<https://github.com/GeomScale/dingo>

From a stoichiometric matrix to a constraint-based model

Reactions

	R ₁	R ₂	R ₃	R ₄	R ₅
Metabolites	-1	0	0	0	0
▲	1	-1	0	0	0
■	0	1	-1	0	0
★	0	1	0	0	-1
●	0	0	1	0	0
◆	0	0	0	-1	0
○	0	0	0	1	-1
◆	0	0	0	0	1

S-matrix

Flux vector

Flux Balance Analysis

Maximize minimize an
objective function:

$$\psi = c_1 v_1 + c_2 v_2 + \dots + c_5 v_5$$

such that:

$$S * v = 0$$

and for each reaction i :

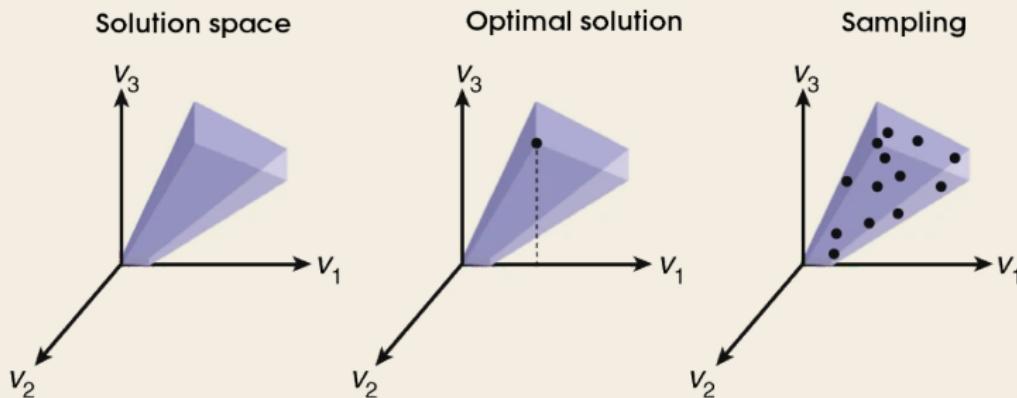
$$lb_i \leq v_i \leq ub_i$$

where lb : lower bound,
 ub : upper bound and

S: the stoichiometric matrix

Flux sampling

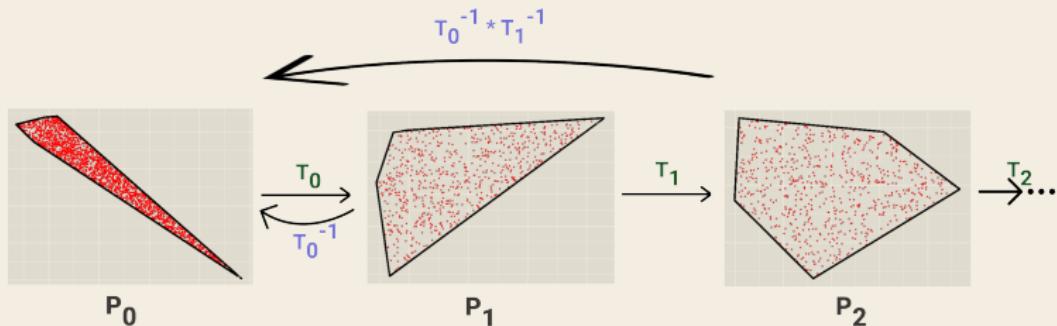
an alternative approach



- enables the analysis of GEMs without the need of an objective function
- determines the feasible solution spaces for fluxes in a network based on a set of conditions as well as the probability of obtaining a solution

Figure from: Heirendt et al. Nature protocols 14.3 (2019): 639-702.

Our Markov Chain Monte Carlo (MCMC) algorithm for flux sampling



Steps of an MMCS phase

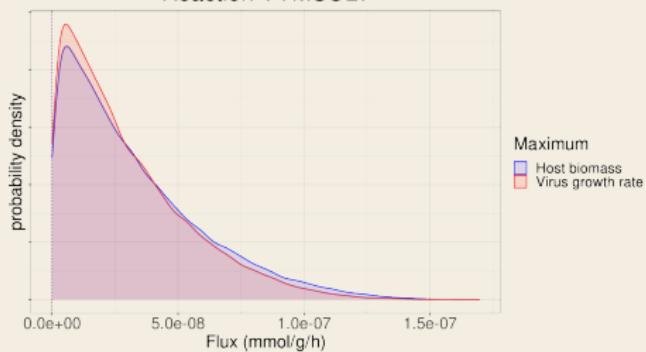
- **sampling step:** using a variant of the **Billiard walk**
- **rounding step:** calculate a linear transformation T_i that puts the sample into isotropic position and then apply it on P_i to obtain the polytope of the next phase
- check several statistic tests

Find possible targets against SARS-CoV-2

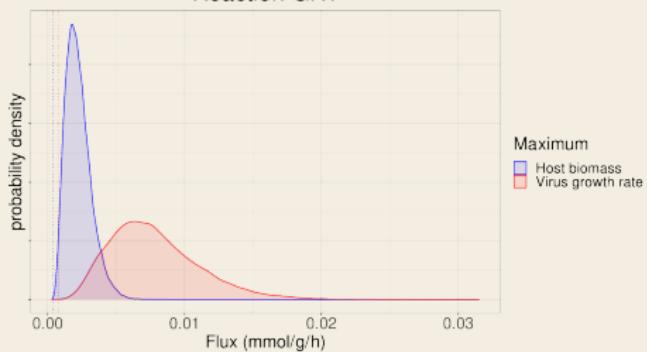
a flux sampling application



Reaction TYMSULT



Reaction GK1

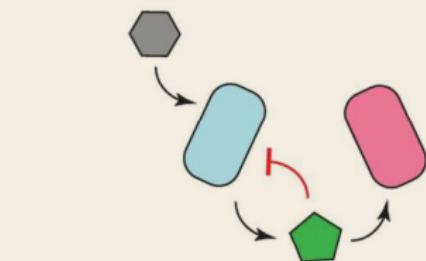


- Check if the flux distribution of a reaction changes.
- Find possible anti-viral targets and study further.

Further applications of metabolic flux sampling



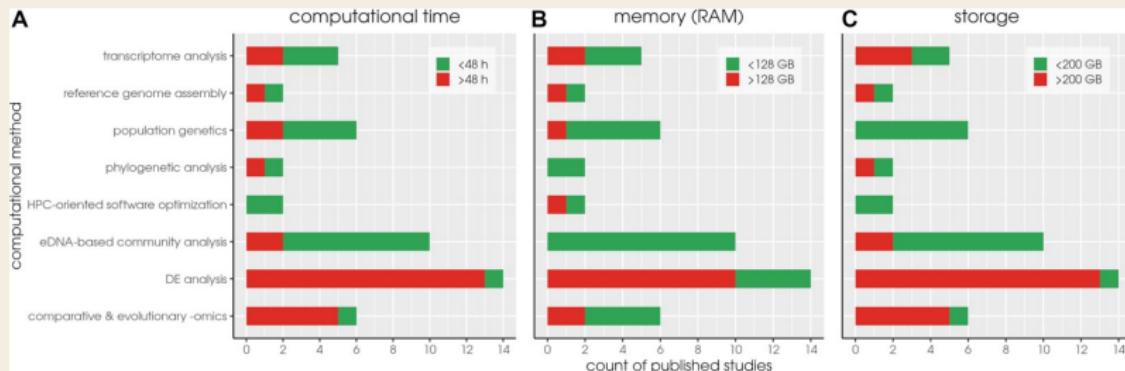
Scott, William T., et al. "Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts." Microbial cell factories 20.1 (2021): 1-15.



What about microbial interactions ?

1. Microbial ecology: a short introduction
2. Bioinformatics methods for microbial diversity assessment
 - 2.1 pema: a metabarcoding pipeline
 - 2.2 darn: known unknowns in COI amplicon data
3. PREGO: a knowledge-base for organisms - environments - processes associations
4. dingo: a Python library for metabolic flux sampling
5. Os & 1s in molecular biology
6. microbetag: enhancing microbial interactions inference from co-occurrence networks
7. Publications

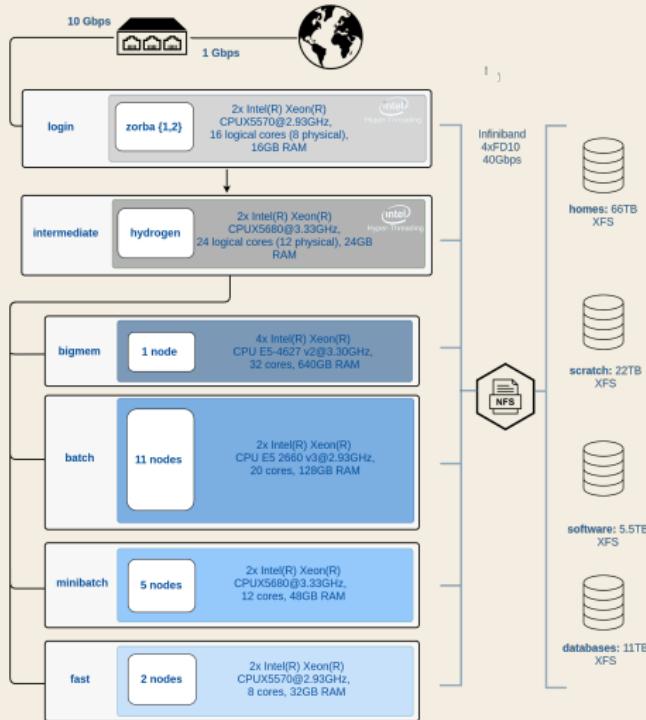
Computational requirements for trivial bioinformatic tasks



Red bars denote published research with high resource requirements
of the various computational methods employed at the IMBBC HPC facility

Zorbas: the HPC facility of IMBBC

a Tier 2 (regional) HPC facility



Block diagram of the
Zorba architecture

Computing infrastructures an alternative for the most!



We will develop a workflow for the analysis of Genomic Observatories (GOs) data that will allow researchers to deal better with the increasing amount of data

1. Microbial ecology: a short introduction
2. Bioinformatics methods for microbial diversity assessment
 - 2.1 pema: a metabarcoding pipeline
 - 2.2 darn: known unknowns in COI amplicon data
3. PREGO: a knowledge-base for organisms - environments - processes associations
4. dingo: a Python library for metabolic flux sampling
5. Os & 1s in molecular biology
6. microbetag: enhancing microbial interactions inference from co-occurrence networks
7. Publications

microbetag: annotating co-occurrence networks

a short term EMBO fellowship

Our short mission consists of 3 modules:

- pathway complementarity
- environmental conditions and phenotypic data integration
- flux sampling on pairs of metabolic models (if possible)

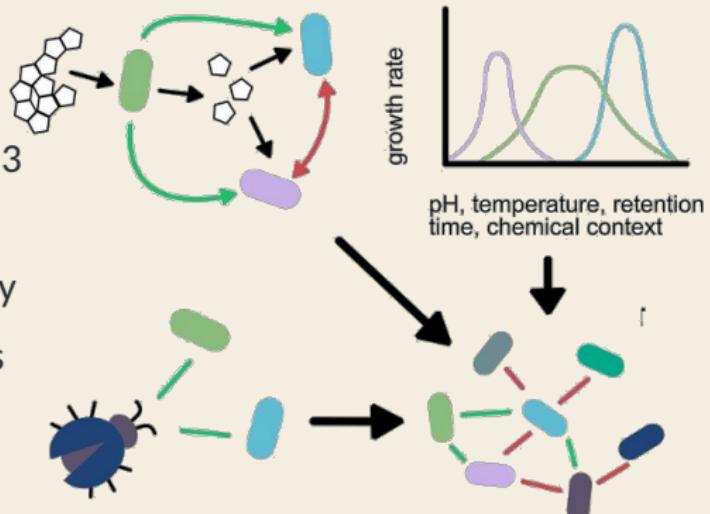


Figure from Röttjers & Faust (2018). FEMS microbiology reviews. 42. 10.1093/femsre/fuy030.

1. Microbial ecology: a short introduction
2. Bioinformatics methods for microbial diversity assessment
 - 2.1 pema: a metabarcoding pipeline
 - 2.2 darn: known unknowns in COI amplicon data
3. PREGO: a knowledge-base for organisms - environments - processes associations
4. dingo: a Python library for metabolic flux sampling
5. Os & 1s in molecular biology
6. microbetag: enhancing microbial interactions inference from co-occurrence networks
7. Publications

Publications

TeX, L^AT_EX, and Beamer

- [1] Zafeiropoulos, H., Paragkamian, S., Ninidakis, S., Pavlopoulos, G.A., Jensen, L.J. & Pafilis, E. PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types. - *under review*
- [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). Os and 1s in marine molecular research: a regional HPC perspective. *GigaScience*, 10(8), giab053.
- [5] Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., ... & 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.

Thank you for your attention
and your patience ;)

GitHub : <https://github.com/hariszaf>

email : haris-zaf@hcmr.gr

Twitter : haris_zaf

web-site : <https://hariszaf.github.io/>



GEOMSCALE