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

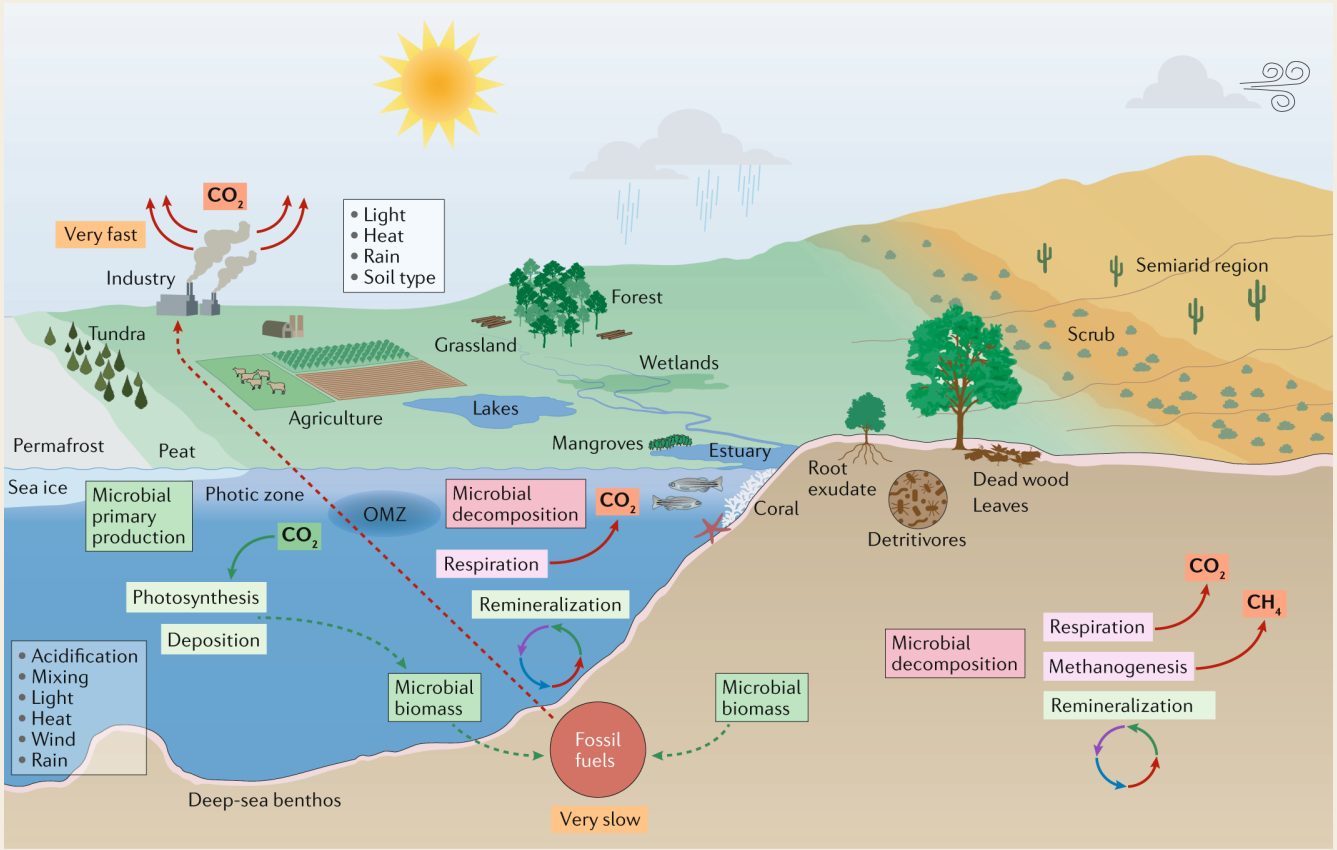


Figure from: Cavicchioli et al. Nature Reviews Microbiology 17.9 (2019): 569-586.

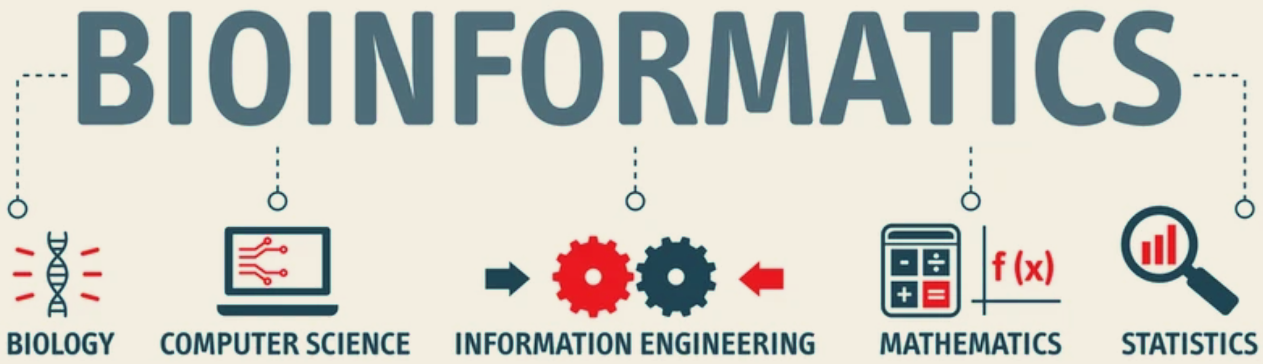
Questions to address

for a deeper understanding of microbial assemblages

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

We are living in a computational era

both a challenge & an opportunity



eDNA metabarcoding

for biodiversity assessment

Marker genes

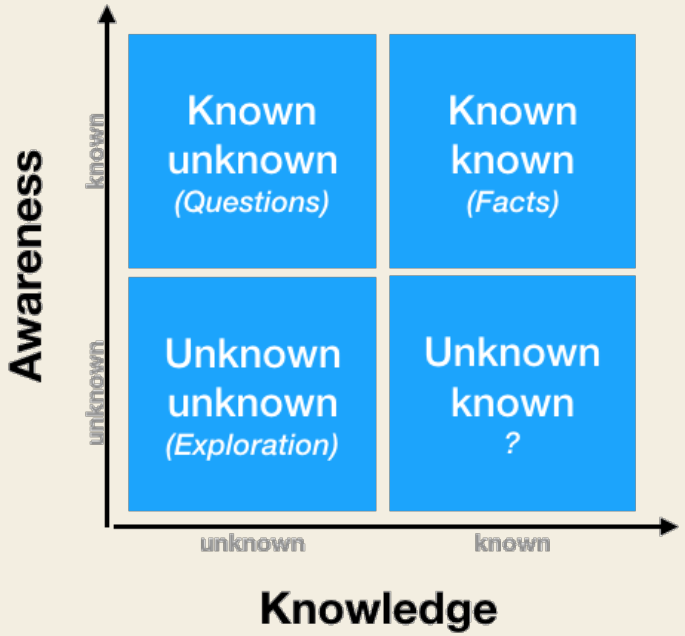
1. **16S rRNA:** Bacteria, Archaea
2. **12S rRNA:** Vertebrates
3. **18S rRNA:** Small eukaryotes, Metazoa
4. **ITS:** Fungi
5. **COI:** Eukaryotes
6. ***rbcl*:** Plants
7. ***dsrb*:** Bacteria, Archaea
8. ...

Methodology

- Sampling
- Extraction
- Bioinformatics
- Biodiversity analysis

Bioinformatics challenges

for the analysis and the interpretation of amplicon data



PEMA

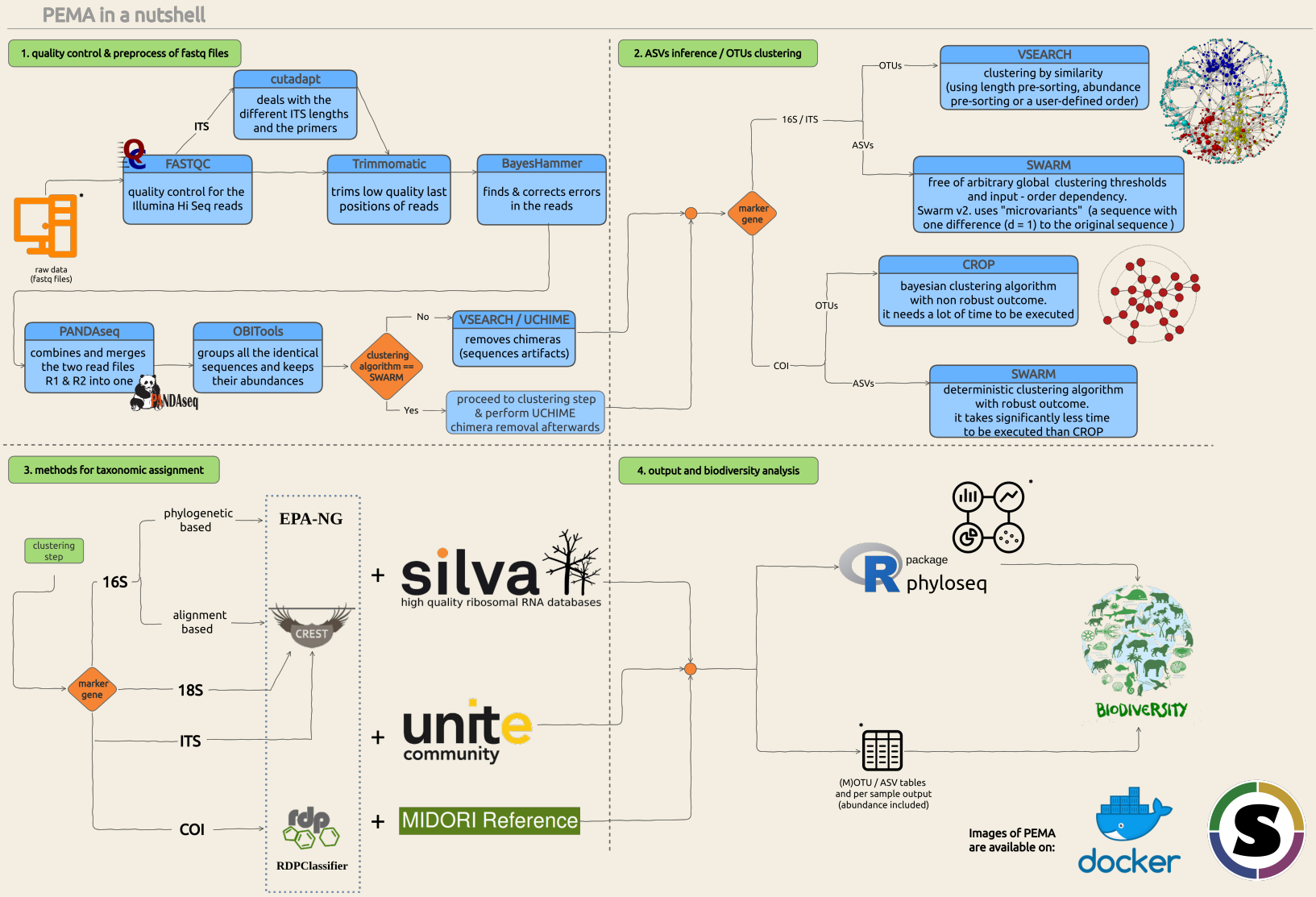
a pipeline for eDNA metabarcoding analysis

https://github.com/hariszaf/pema

pema.hcmr.gr

PEMA features

one step at a time!



* Illustrations with asterisk are from the Noun Project

PEMA coding insights

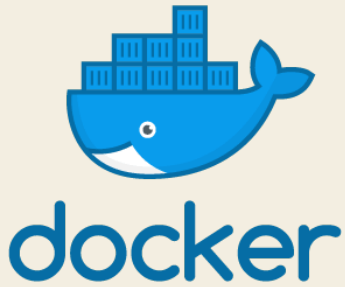
Being a geek just for a bit !

```
for(int i : range(1,
    in  := "in_$i.tx
    sys date > $in

    out := "out_$i.t
    task( out <- in
        sys echo Tas
    }
}
```

`BigDataScript`

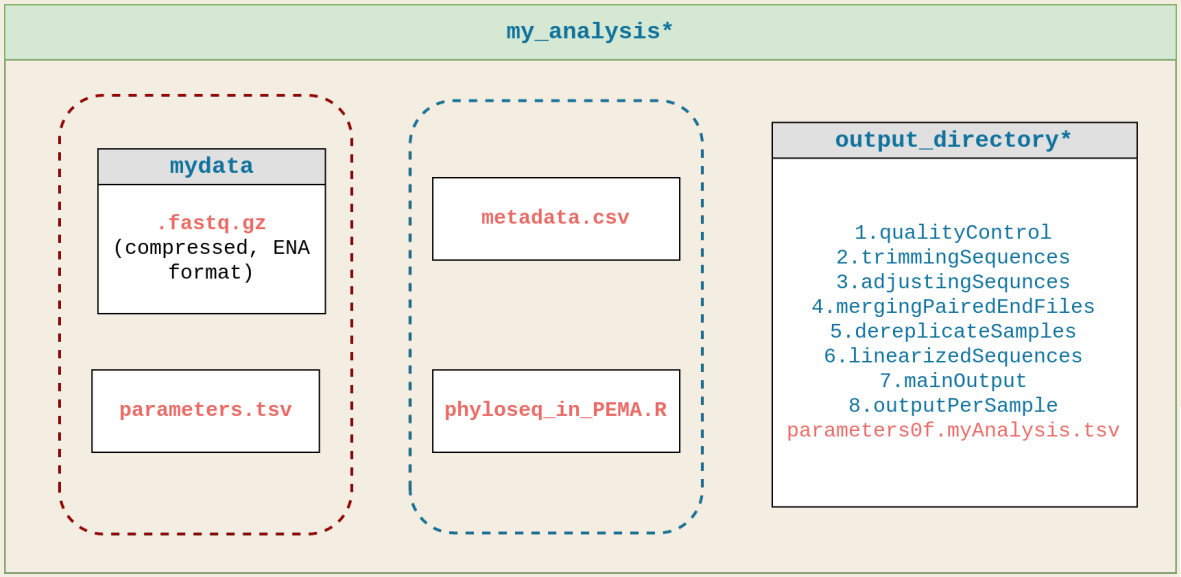
programming language



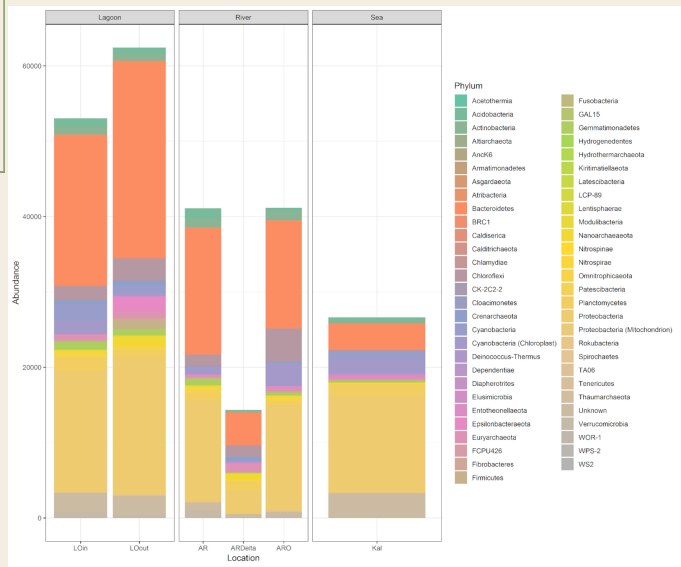
Containerization

Mount your I/O

give & take



	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

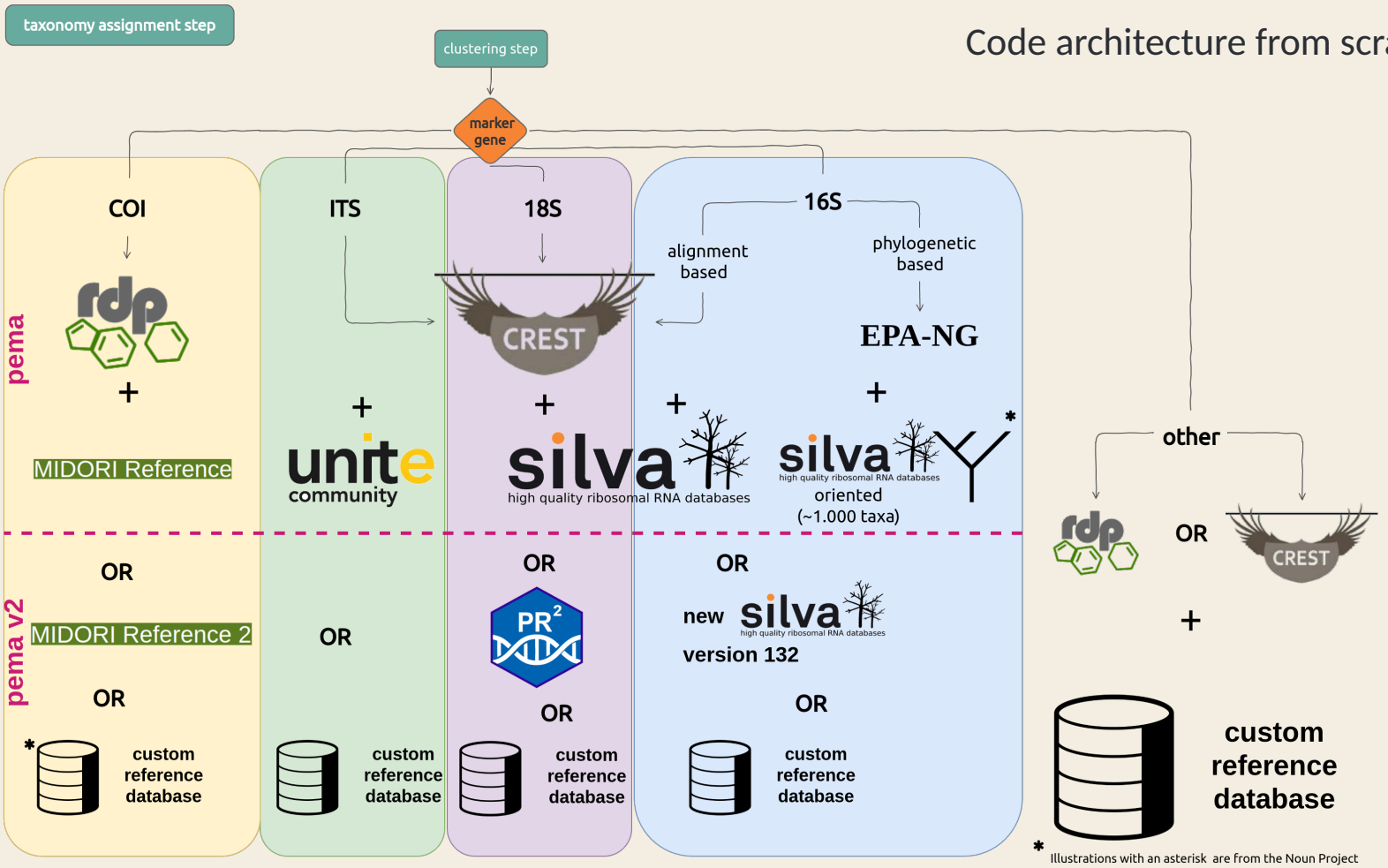


text file
directory
***user can edit the name**
(the rest **need** or will have the exact names shown)

- - - - - mandatory input files
- - - - - optional input files

PEMA v.2

addressing some of the challenges



Code architecture from scratch!

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

darn

https://github.com/hariszaf/darn

Dark mAtteR iNvesigator

investigating known unknown in COI amplicon data

What is all these unassigned OTUs / ASVs?

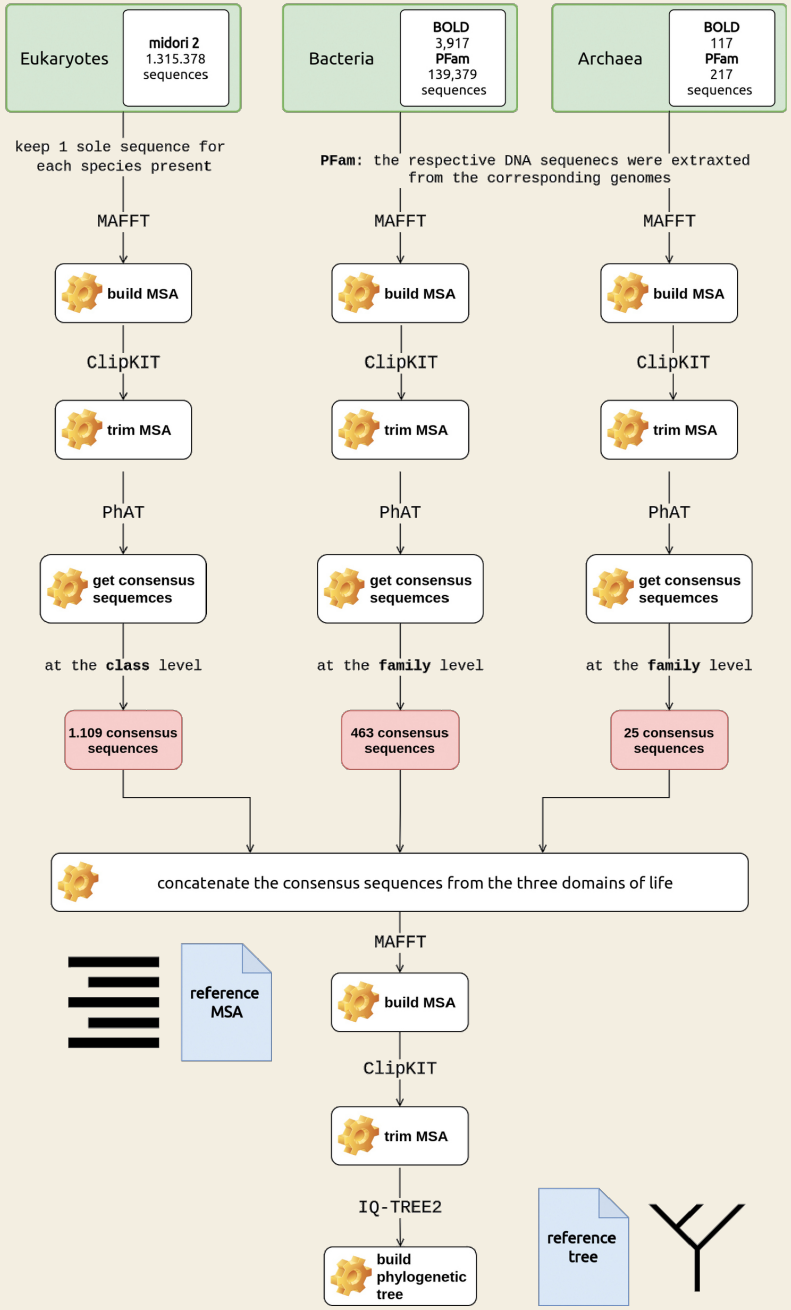
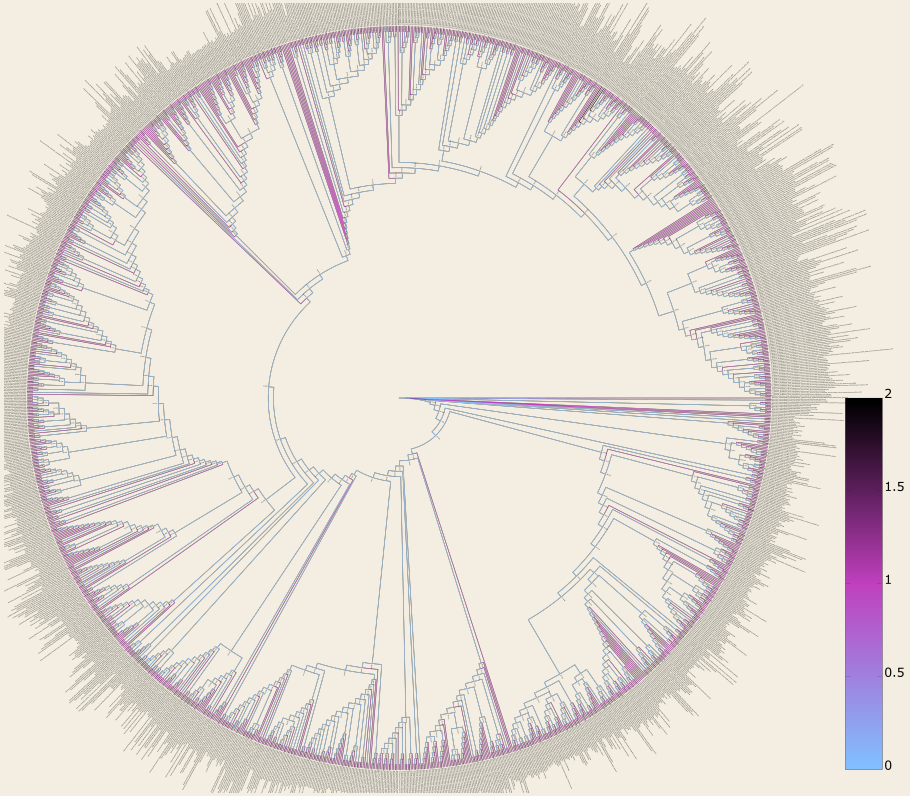


Figure from: 10.3897/mbmg.5.69657

Phylogeny of the COI consensus sequences retrieved

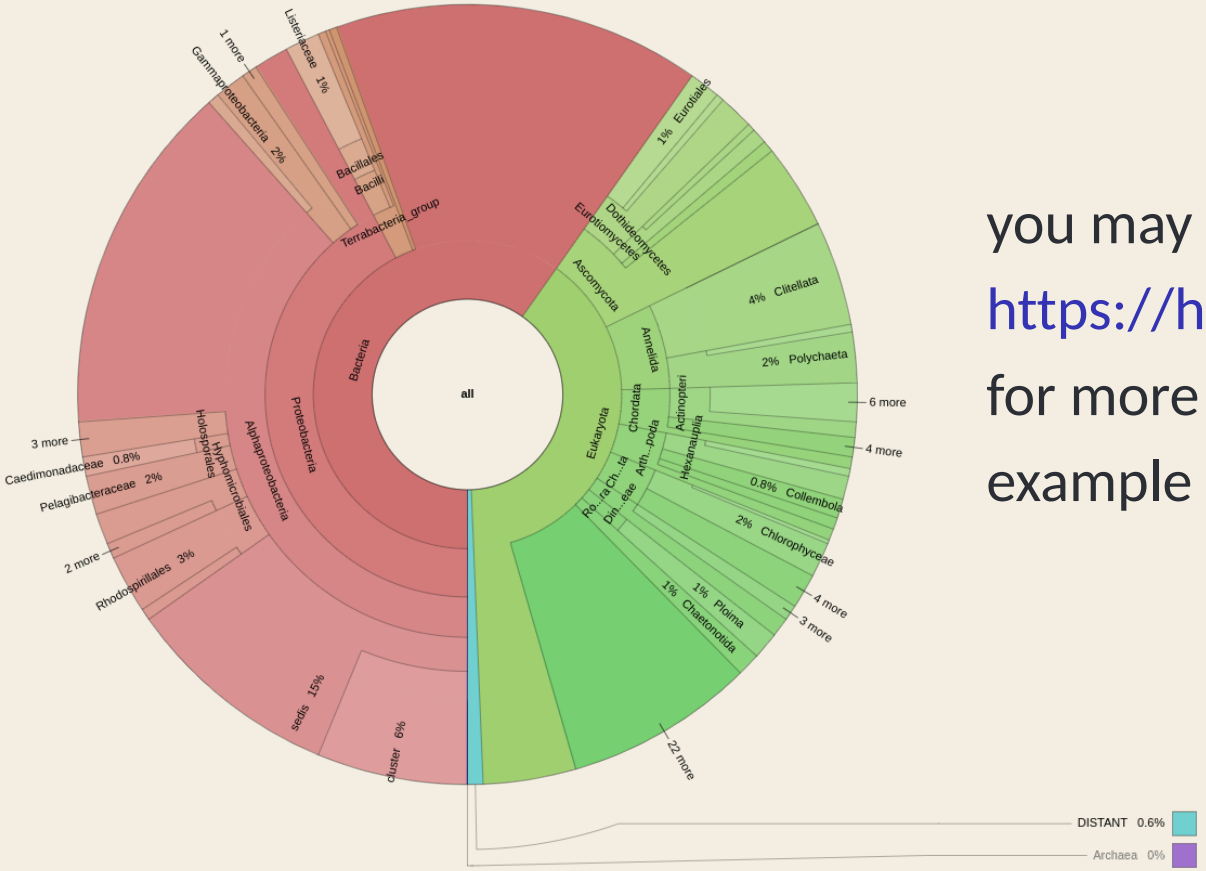
the tree that DARN makes use of



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/

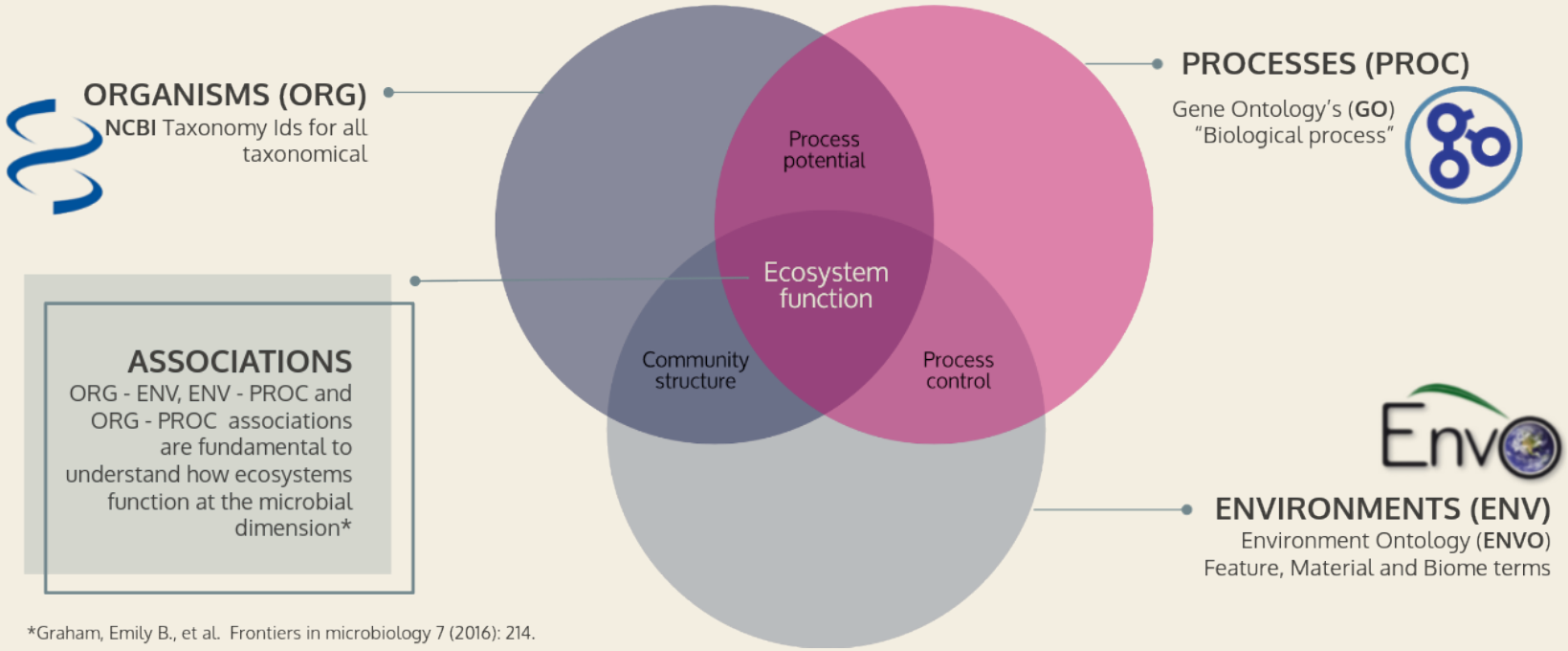
for more DARN output example cases

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 [-]

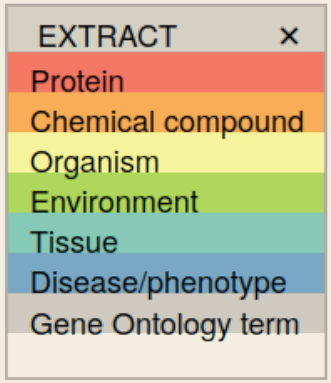
Collection date:	11/1/11
Elevation:	200
Environment (biome):	soil
Environment (feature):	nosZ
Environment (material):	soil DNA
Environmental package:	MIGS/MIMS/MIMARKS.soil
Geographic location (depth):	15-20cm
Instrument model:	454 GS FLX Titanium
Investigation type:	miens-survey
NCBI sample classification: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

Named Entity Recognition

tagging the literature



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.

PREGO methodology

co-occurrence again!

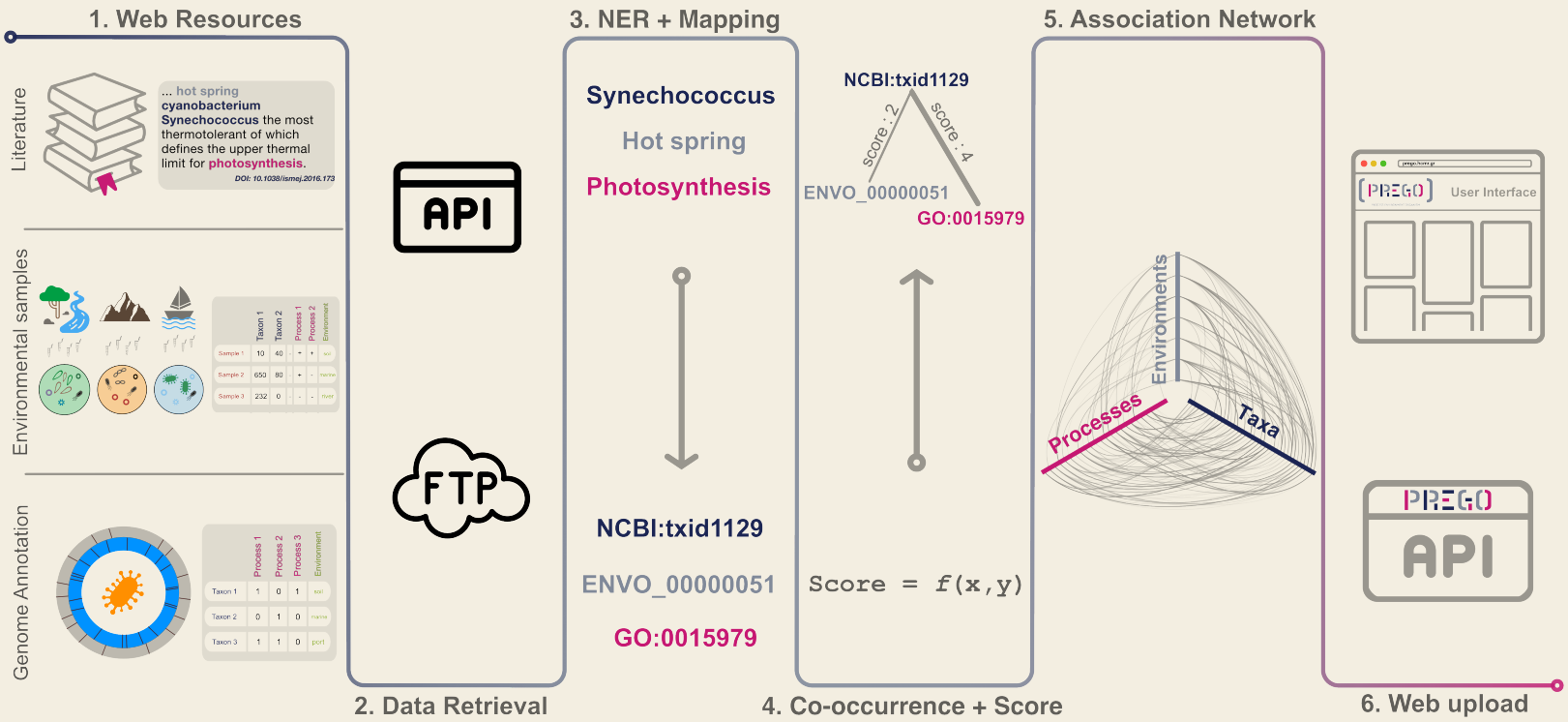
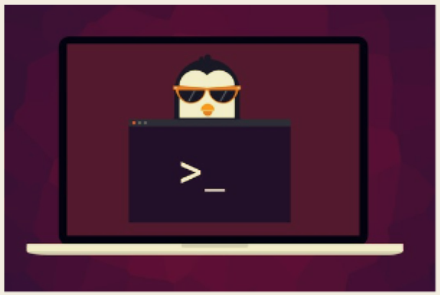


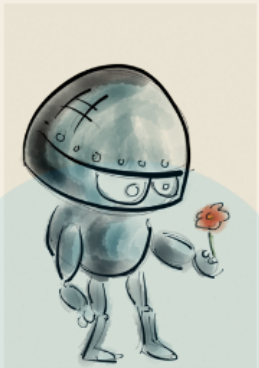
Figure from the PREGO publication that is now under review.

Building a knowledge-base

development and information technology operations

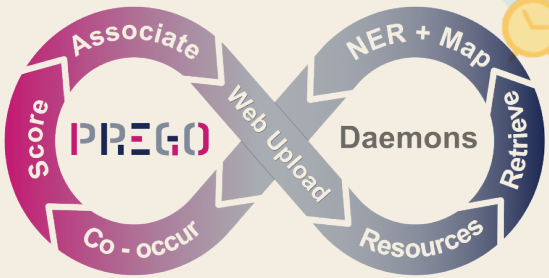


frequently updates;
`robots` allow to keep up with
literature and metagenomic data



building a
knowledge base (KB)

server dedicated
for building and
accessing the KB



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-littoral zone	1.4	★☆☆☆☆
Oil spill	1.4	★☆☆☆☆
Petroleum	1.1	★☆☆☆☆
Sea floor	1.1	★☆☆☆☆

Showing 1 to 12 of 12 entries

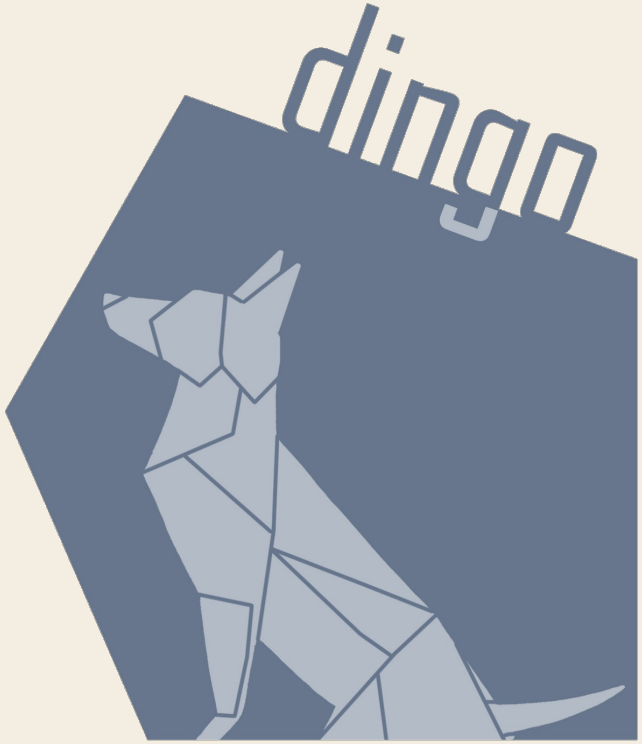
Environments | Biological Processes | Molecular Function | Documents | Downloads

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	★☆☆☆☆

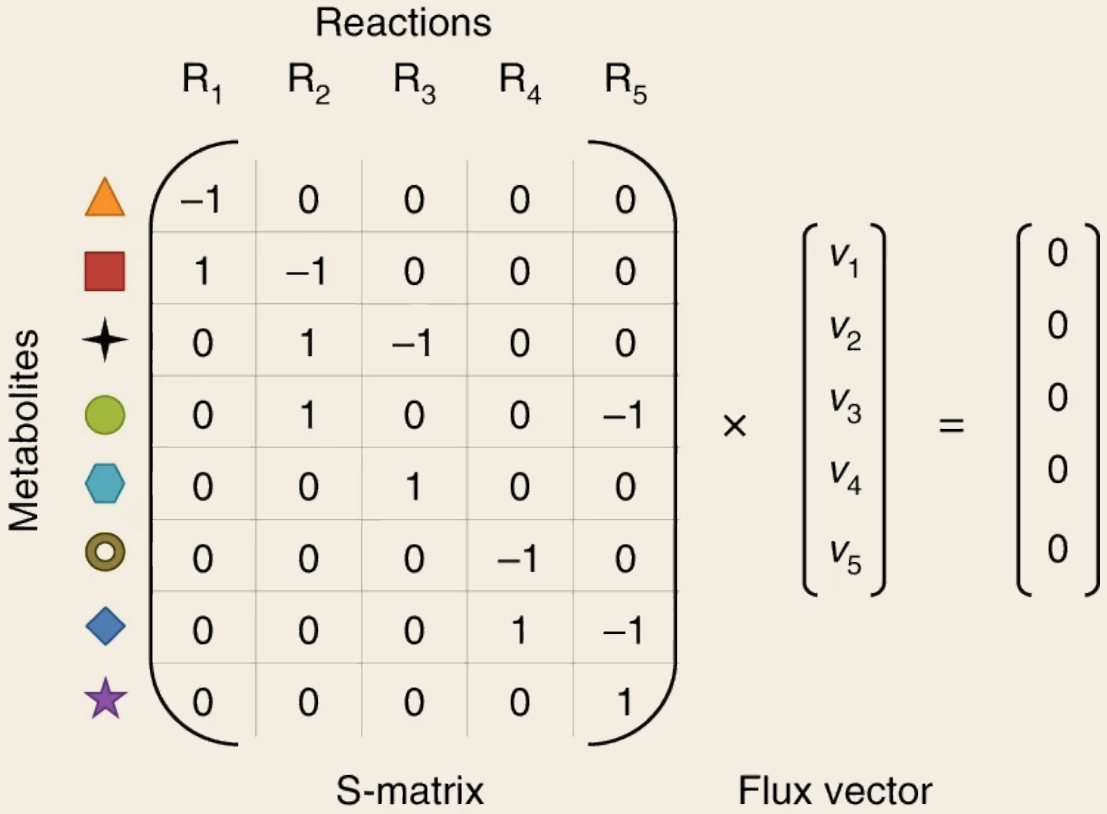
Showing 1 to 14 of 14 entries



https://github.com/GeomScale/dingo

From a stoichiometric matrix

to a constraint-based model



Flux Balance Analysis

Maximize minimize an objective function:

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

such that:

$$S * v = O$$

and for each reaction i:

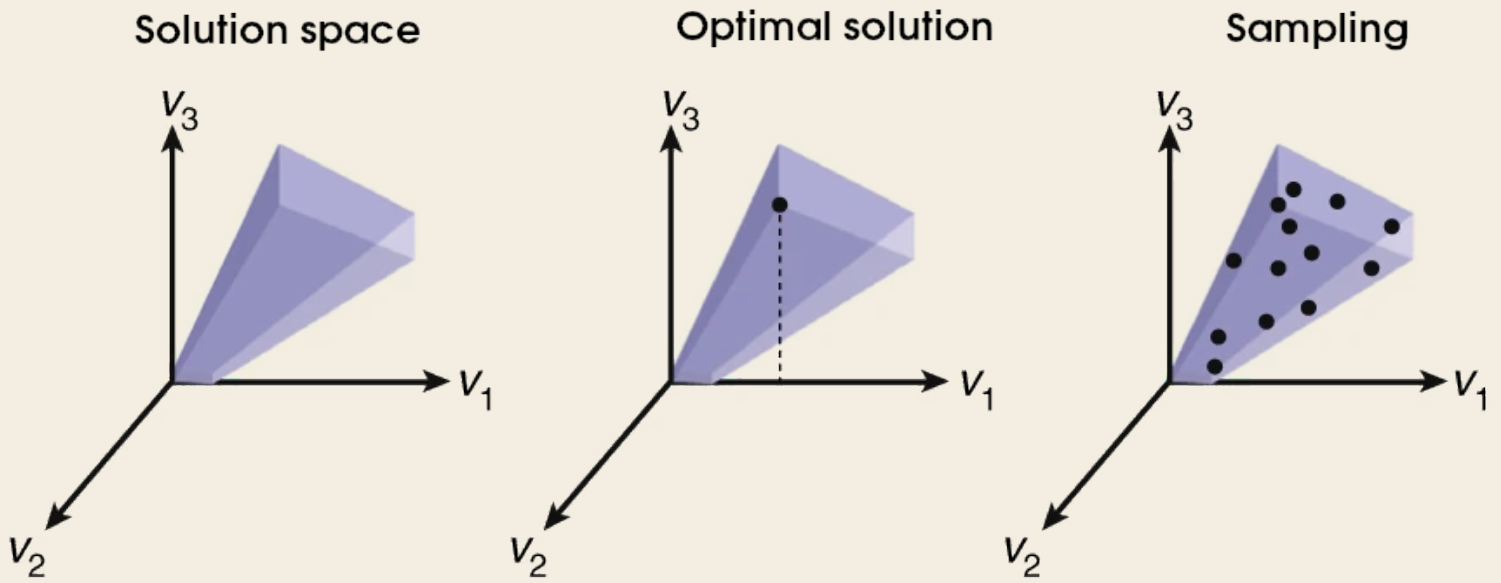
$$lb_i <= v_i <= ub_i$$

where lb: lower bound, ub: upper bound and S: the stoichiometric matrix

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

Flux sampling

an alternative approach

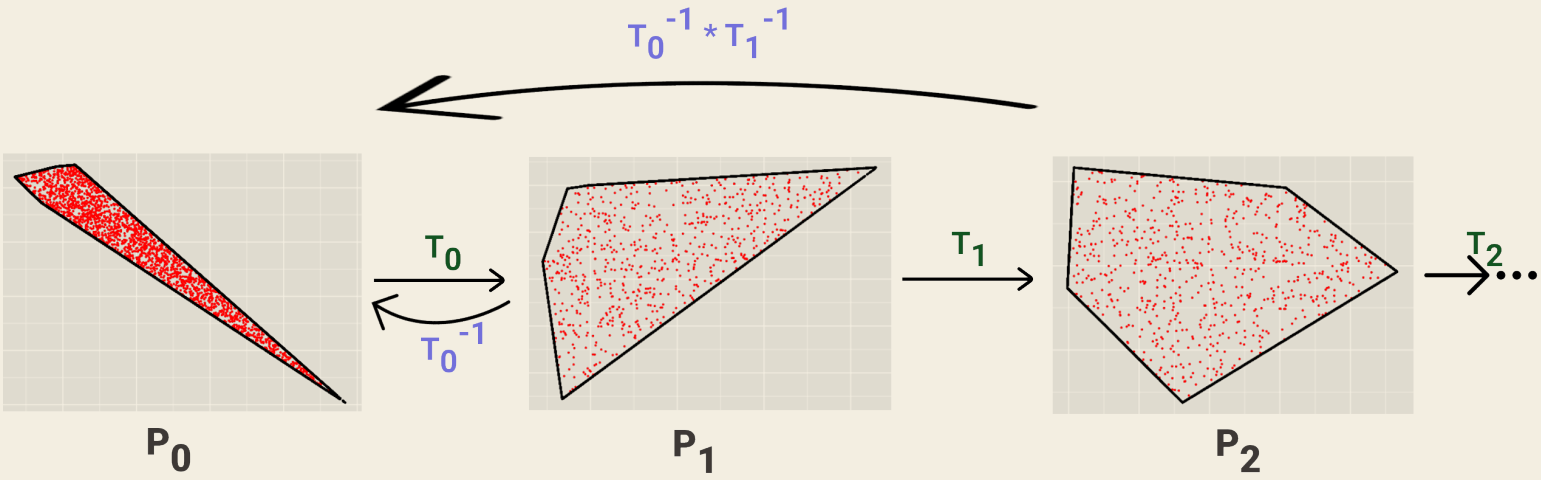


- enables the analysis of GEMs without the need of an objective function
- determines the feasible solution spaces for fluxes in a network based on a set of conditions as well as the probability of obtaining a solution

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

Our Markov Chain Monte Carlo (MCMC) algorithm

for flux sampling

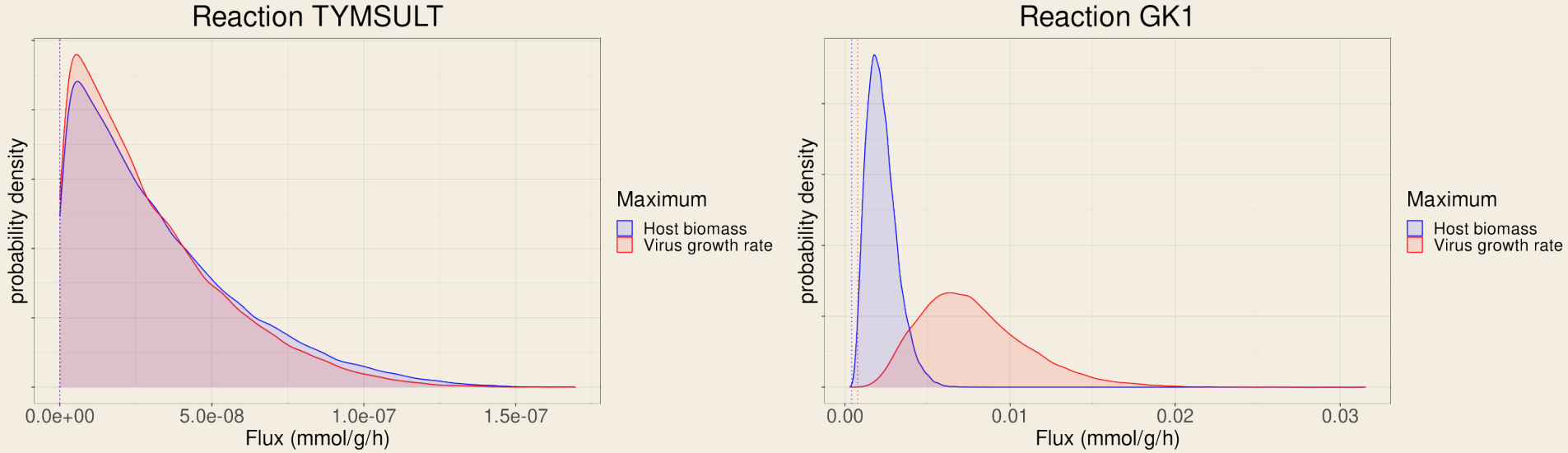


Steps of an MMCS phase

- **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

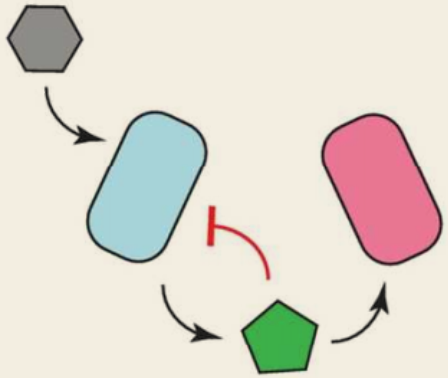


- Check if the flux distribution of a reaction changes.
- Find possible anti-viral targets and study further.

Further applications

of metabolic flux sampling

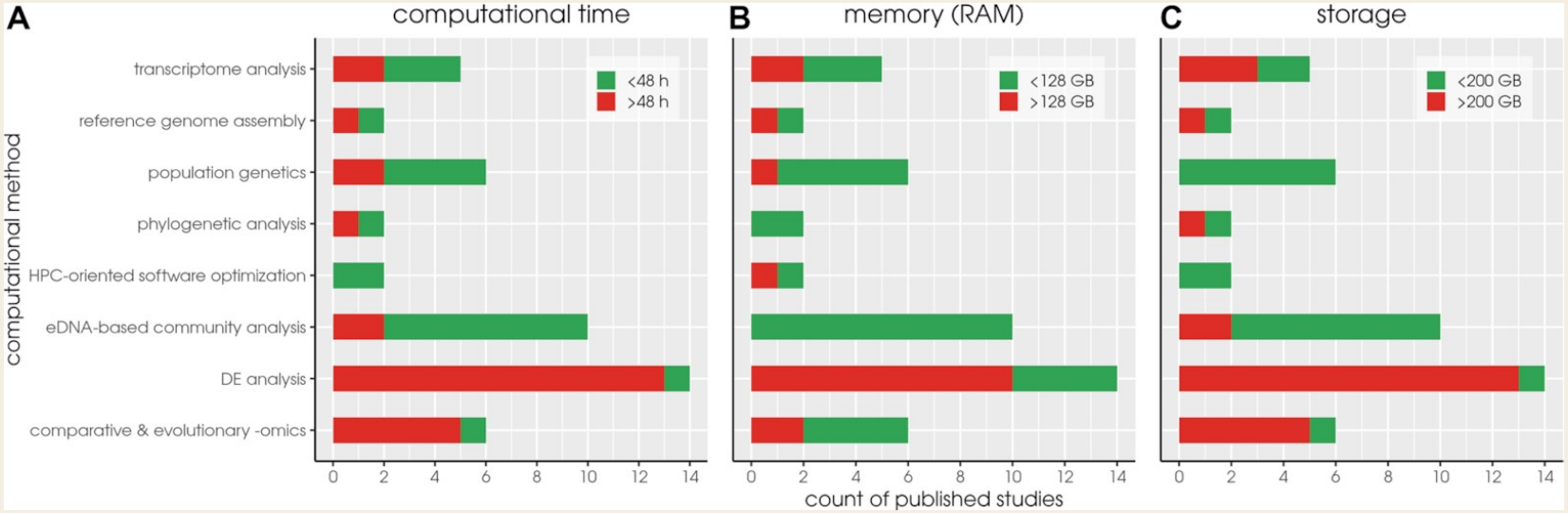
Scott, William T., et al. "Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts." Microbial cell factories 20.1 (2021): 1-15.



What about microbial interactions ?

Computational requirements

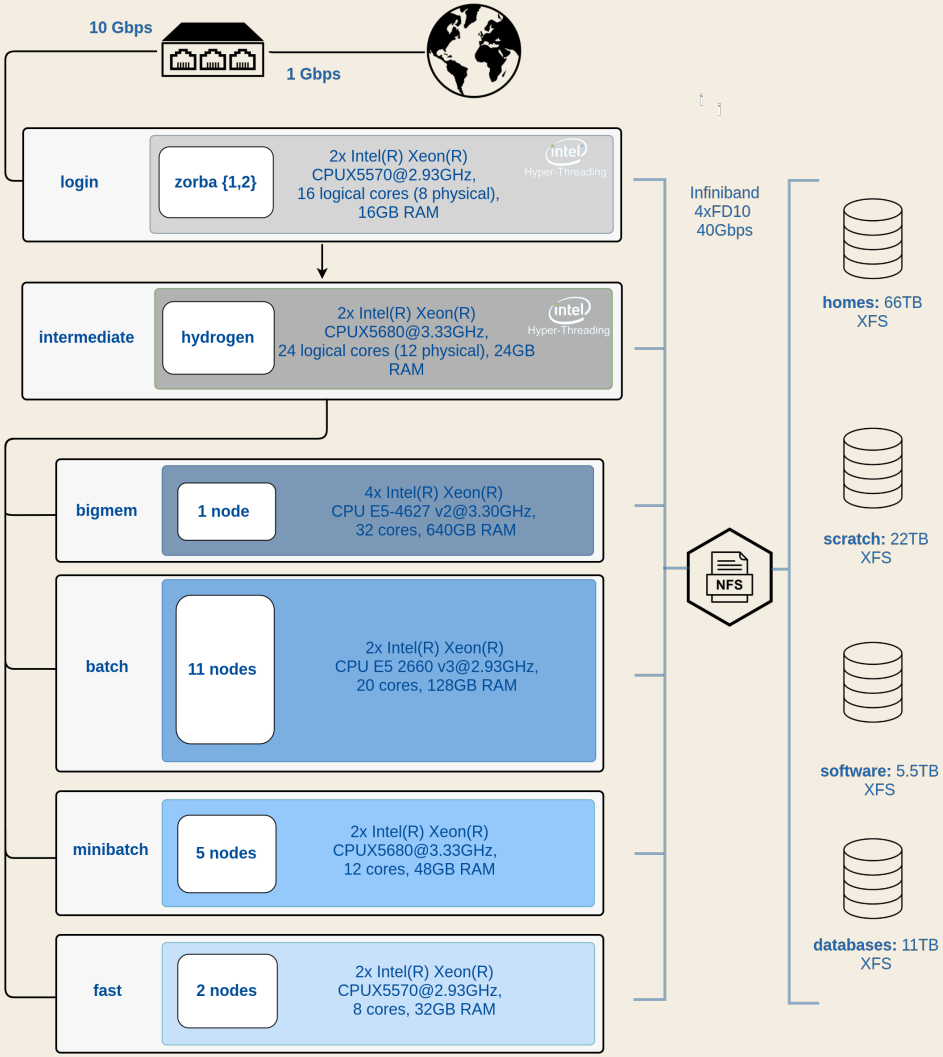
for trivial bioinformatic tasks



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

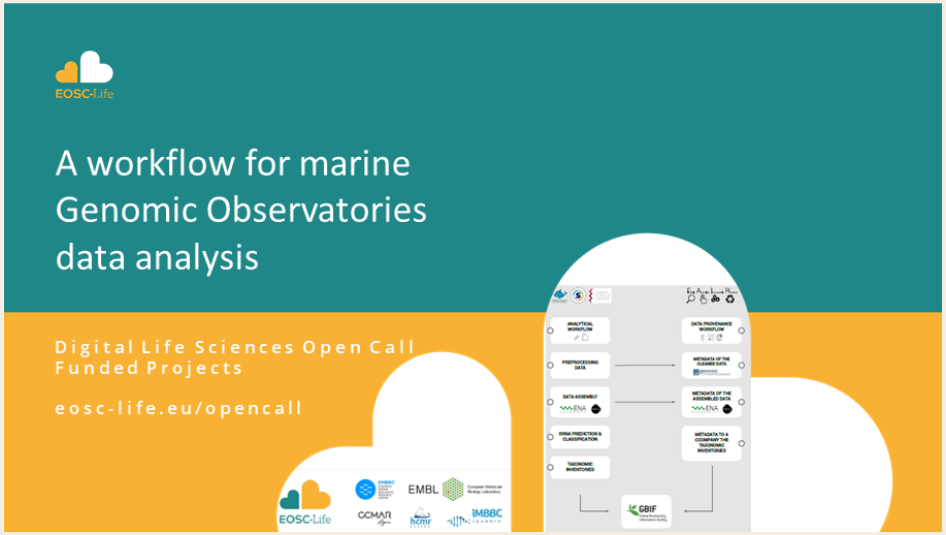
a Tier 2 (regional) HPC facility



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

`microbetag`: annotating co-occurrence networks

a short term EMBO fellowship

Our short misssion consists of 3 modules:

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

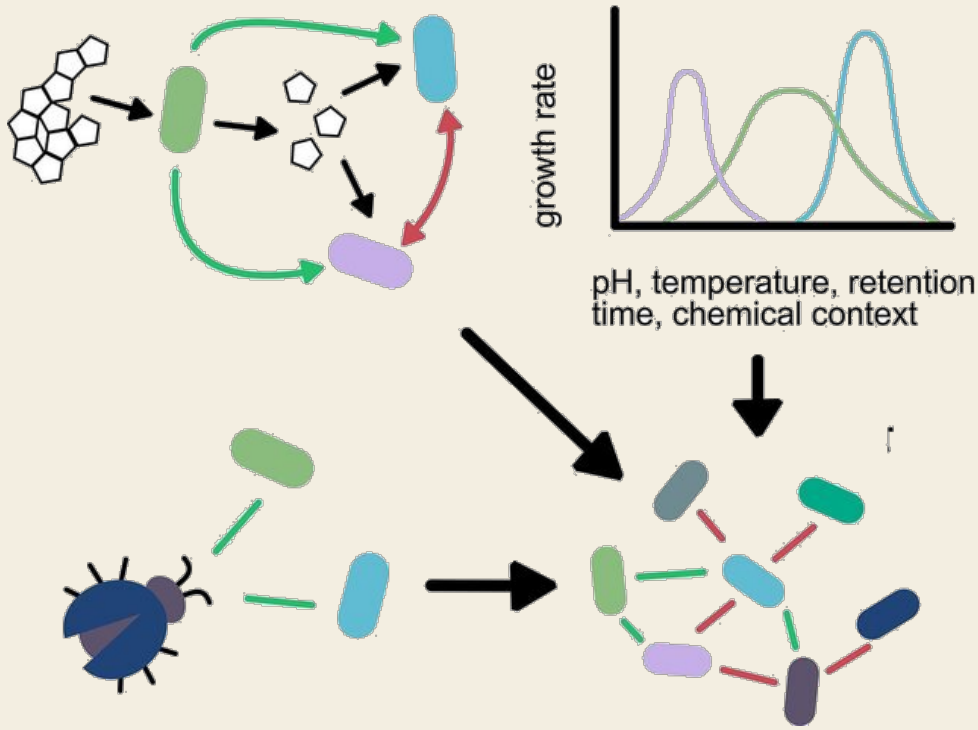


Figure from Röttjers & Faust (2018). FEMS microbiology reviews. 42. 10.1093/femsre/fuy030.

First steps

pathway complementarity

1. use abundance & metadata table to run FlashWeave
2. get the NCBI Taxon id of each taxon present in the edge file
3. search for available reference genomes for these ids
4. use KEGG modules to get major metabolic pathways of interest
5. search for pathway complementarity in each module of interest

Relative work

- Levy & Borenstein. Proceedings of the National Academy of Sciences 110.31 (2013): 12804-12809.
- Zelezniak et al. Proceedings of the National Academy of Sciences 112.20 (2015): 6449-6454.

Databases to exploit

and how to

To get KO and further annotations of a taxon:

- via KEGG organisms using KEGG API
- via JGI using the `ecg` tool
- via BacDive using the BacDive API

Further annotation sources to be considered:

- FAPROTAX
- BugBase
- SigMol

Open questions

just a few of them ;)

- species - strain inheritance in data integration
- dataset, maybe the Venturelli one
- metabolic modelling at the community level
- competition and mutualism: a dialectic relationship

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. 0s & 1s in molecular biology
6. `microbetag`: enhancing microbial interactions inference from co-occurrence networks
7. Publications

Publications

TEX, LATEX, 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). 0s 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/

Spacial thanks to:

Dr. Pafilis E.

Dr. Pavloudi C.

PhD Paragkamian S.

Dr. Chalkis A.

Prof. Tsigaridas E.

Dr. Fisikopoulos V.

GeomScale